Higher-Order Numerical Schemes and Operator Splitting for Solving 3D Paraxial Wave Equations in Heterogeneous Media

Eliane Bécache, Francis Collino, Patrick Joly

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Higher-order numerical schemes and operator splitting for solving 3D paraxial wave equations in heterogeneous media

Eliane BÉCACHE, Francis COLLINO, Patrick JOLY

N° 3497

Septembre 1998

THÈME 4

Rapport de recherche

ISSN 0249-6399

INRIA
ROCQUENCOURT

Higher-order numerical schemes and operator splitting for solving 3D paraxial wave equations in heterogeneous media

Eliane BÉCACHE*, Francis COLLINO†, Patrick JOLY‡

Thème 4 — Simulation et optimisation
de systèmes complexes
Projet Ondes

Rapport de recherche n ° 3497 — Septembre 1998 — 42 pages

Abstract: We investigate numerical schemes for solving 3D paraxial wave equations that are compatible with the use of splitting methods without losing accuracy. The novelty of these paraxial equations (introduced in [14]) compared with classical alternate directions methods is to use more than the two usual cross-line and in-line directions for the splitting. It gives rise to a series of 2D extrapolations in each direction of splitting. Propagation along depth is done with a higher-order method based on a conservative Runge Kutta method. The discretization along the lateral variable is done using higher-order finite difference variational schemes. We present a detailed plane wave analysis in a homogeneous medium that leads to a classification of several particular schemes with respect to the numerical dispersion they generate. The dispersion analysis extended to 3D helps chosing the "best" coefficients of the extrapolation operators on the dispersion point of view. We conclude with numerical experiments in 2D as well as in 3D homogeneous and heterogeneous media.

Key-words: seismic migration, paraxial wave equations, splitting, higher-order finite differences, finite elements, numerical dispersion.

(Résumé : tsvp)

* INRIA-rocquencourt. Eliane.Becache@inria.fr
† INRIA-rocquencourt. Francis.Collino@inria.fr
‡ INRIA-rocquencourt. Patrick.Joly@inria.fr

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
Téléphone : 01 39 63 55 11 - International : +33 1 39 63 55 11
Télécopie : (33) 01 39 63 53 30 - International : +33 1 39 63 53 30

Schémas d'ordre élevé et splitting d'opérateurs pour la résolution d'équations paraxiales en milieu hétérogène

Résumé : Nous proposons des schémas numériques pour la résolution d'équations paraxiales 3D qui sont compatibles avec l'utilisation de méthodes de splitting, sans perte de précision. La nouveauté de ces équations paraxiales (introduites dans [14]), en comparaison avec des méthodes classiques de directions alternées, est qu'elles utilisent plus de directions de splitting que les deux habituelles (suivant les axes). Ces équations conduisent à la résolution d'une série d'extrapolations 2D dans chaque direction du splitting. La discrétisation suivant la profondeur se fait grâce à une méthode d'ordre élevé basée sur une méthode conservative de Runge Kutta. En ce qui concerne la discrétisation par rapport à la variable latérale, nous proposons une méthode d'ordre élevé de différences finies variationnelles. Nous présentons una analyse détaillée des ondes planes en milieu homogène qui conduit à una classification de plusieurs schémas particuliers par rapport à leur dispersion. L'analyse de dispersion étendue au 3D permet de choisir les "meilleurs" coefficients des opérateurs d'extrapolation du point de vue de leur dispersion. Nous présentons pour finir des résultats numériques aussi bien en 2D qu'en 3D, en milieux homogène et hétérogène.

Mots-clé : migration sismique, équations paraxiales, ondes, splitting, différences finies d'ordre élevé, éléments finis, dispersion numérique.

Table of Contents

1 Introduction

Paraxial approximations of the wave equation are commonly used in many applications when the waves propagate near a privileged direction. Although some of them have been especially designed to handle time-domain problems (see for instance [15], [34]) we will focus here on the most common frequency-domain approach : it starts with a Fourier decomposition of the source function, then transforms the linear hyperbolic time-domain wave equation into an elliptic Helmholtz equation and finally solves the propagation problem for each frequency. For waves propagating near a particular direction, the Helmholtz equation can be seen as the factorization of two non-local integro-differential one-way wave equations, for which this privileged direction plays the role of the evolution variable. The square root operator involved in these two one-way wave equations can then be approximated in several ways leading to local parabolic partial differential equations. The paraxial approximation is valid as long as the wave remains “close enough” to the privileged direction, more precisely the accuracy depends on the order of magnitude of the error between the exact square root and its approximation and of course the better is the approximation the wider is the allowed propagation angle. That is how several families of approximation have been designed with the denomination of fifteen, forty-five or sixty degree approximations. For readers interested in the mathematical aspects of the paraxial equations, we refer to the paper by Collino and Joly [14] and to its bibliography.

One of the main application of paraxial approximations is the resolution of range-dependent ocean acoustic propagation problems, and the range r becomes the evolution variable. Tappert's parabolic equation [37] was the first paraxial approximation introduced to the underwater acoustics community and numerous contributions have been made since then (see D. Lee and A. D. Pierce [28]). The other application, the one we have in mind in this paper, is the migration in geophysics. This time, the evolution variable is the depth z. The paraxial approximation is used in this case for solving the downward extrapolation in the subground of a wave-field known at the surface. This is actually one step in the solution of the full inverse problem. In this area also, these equations have been extensively applied. In particular Claerbout [11] was the first to introduce fifteen degree and forty-five degree type equations for the extrapolation of 2-D seismic data.

Concerning the discretization of these equations, two different kinds of variables appear naturally : the depth variable z and the lateral variables x_1 for 2D problems, (x_1, x_2) for 3D problems. There are essentially two different approaches for solving these equations. The first one is based on the use of discrete extrapolation operators (see for instance [22, 21, 23]). The second approach consists in approximating the equation with Finite Differences or Finite Elements. In this case the discretization is done using a Crank-Nicolson scheme in the depth variable and a second order finite difference scheme in the lateral variables. Although these equations are commonly used for solving 2D problems (eg Brysk [9], Dubrulle [16], Ma [32], Ristow and Rühl [35]) their solution is not so obvious for 3D problems and requires to solve at each extrapolation step a 2D problem in the transverse plane that gives rise to a large linear system difficult to invert. Joly and Kern [24] and Kern [25], [27] have proposed using modern iterative methods to solve it. We propose here to explore another approach. The general outlines of these two alternative approaches are presented in [6].

To avoid the resolution of a transverse 2D problem, Collino and Joly [14] have constructed new families of paraxial approximations that are compatible with splitting methods. The novelty of their approach in comparison with classical alternate direction methods ([8, 19, 29, 18]) is to introduce other directions for the splitting than the usual cross-line and in-line directions. This allows to get forty-five degree and sixty degree accurate approximations. In their paper, they explain in details how to construct families of approximations with 3, 4 and any given number of directions of splitting, and also with one or more fractions per direction and they make the analysis of their accuracy with the help of Taylor expansions. The undesirable anisotropic effects observed with Brown's approximation [8](alternate directions using only the x_1 and x_2 directions) disappear with the new equations. The problem is then reduced to a series of 2D extrapolations in each direction of splitting. They have presented these new equations for constant velocities, but there is no difficulty to extend them to the case of heterogeneous velocities, following the criteria given by Bamberger and al. [4]. Recently and independently of this work, D. Ristow and T. Rühl [36] used the same idea of operator splitting in alternate directions. They describe three, four and six-way splitting, and propose two different methods for calculating the coefficients of the operators. The first one is, as Collino and Joly, the Taylor expansion method. The second one is based on a constrained optimization procedure.

The aim of this paper is to describe a systematic way to get accurate discretizations, both in depth and in the lateral variables to solve the new 3D paraxial equations with splitting methods. The dispersion analysis is a well known and very efficient tool to estimate the quality of a given numerical scheme. This has been studied for the approximation of 2D paraxial equations in [12]. The numerical dispersion becomes even more important in 3D problems, that is why we propose here some new higher-order numerical schemes that attenuate these effects. We construct our schemes in general heterogeneous media and study their accuracy via a dispersion analysis in homogeneous media.

The paper is organized as follows. Section 2 is devoted to brief recalls about the classical paraxial approximations. In section 3, we set the PDEs associated to the new paraxial approximations and end up with a 2D model paraxial equation. In section 4, we are concerned with the discretization of this 2D equation. Higher-order discretizations in the depth variable are presented in subsection 4.1. They are obtained with the procedure presented by Kern [25] for the full 2D paraxial equations based on a conservative implicit Runge-Kutta method. Subsection 4.2 presents the discretization in the lateral variable with variational finite differences techniques. With these techniques, the variable coefficients are easily taken into account. Section 5 is devoted to the dispersion analysis of these schemes, through the propagation of plane waves in homogeneous media. This analysis allows us to classify the schemes from the dispersion point of view. In section 6 we come back to the 3D paraxial equations: first, in §6.1, with the 3D algorithm; then, in §6.2, with the extension of the dispersion analysis to the family of 3D forty-five degree approximations which gives a new criterion to choose the "best" coefficients for the forty-five degree paraxial equation, "best"in the sense that they minimize the dispersion. Section 7 presents several numerical experiments for which we compare the schemes up to the order 6.

2 The classical approach

This section is devoted to the classical paraxial approximations. A classical class of well-posed higher order approximations is presented. After briefly setting out the splitting methods principle, we describe the way they are classically applied for solving discretized 3D paraxial equations.

2.1 The classical continuous paraxial equations

We consider the propagation in the whole homogeneous space. In the sequel, t denotes time, (x_1, x_2, z) are space variables, z is the privileged direction, $x = (x_1, x_2)$ are the transverse variables. We assume the velocity c to be constant. Classically, the solution of the wave equation in the whole space

$$\frac{1}{c^2}\frac{\partial^2 v}{\partial t^2} - \frac{\partial^2 v}{\partial x_1^2} - \frac{\partial^2 v}{\partial x_2^2} - \frac{\partial^2 v}{\partial z^2} = 0, \tag{1}$$

with appropriate boundary and initial conditions, can be split into two waves, an up-going wave and a down-going wave. The up-going wave, that we are interested in, satisfies the one-way wave equation

$$\frac{d\hat{v}}{dz} + i\frac{\omega}{c}\left(1 - \frac{c^2\,|k|^2}{\omega^2}\right)^{\frac{1}{2}} \hat{v} \;=\; 0, \quad z \geq 0, \tag{2}$$

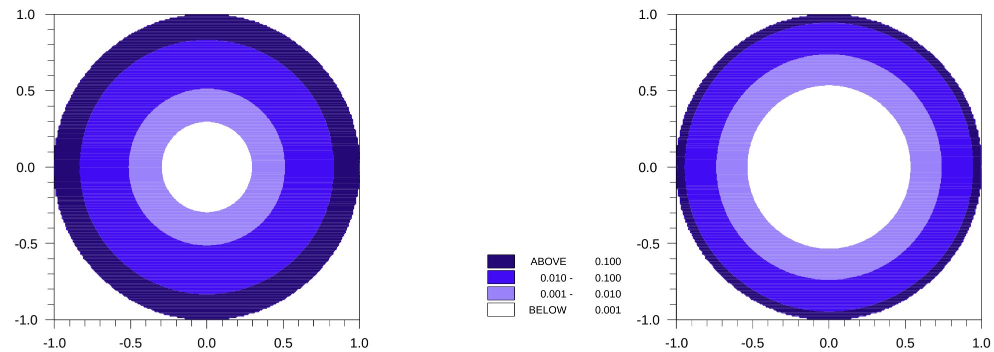
where $\hat{v}$ denotes the Fourier transform of v with respect to t (time) and x_1, x_2

$$\hat{v}(k_1, k_2, z, \omega) = \int\int\int v(x_1, x_2, z, t)e^{i(k_1x_1 + k_2x_2 - \omega t)}dx_1dx_2dt. \tag{3}$$

Paraxial equations are approximations of the one way up-going wave obtained by replacing the square root in (2), $(1- \mid \kappa \mid^2)^{1/2}$ (where $\kappa = (\kappa_1, \kappa_2)$ and $\kappa_1 = \frac{ck_1}{\omega}$, $\kappa_2 = \frac{ck_2}{\omega}$), by rational fractions. We shall denote such approximations by $(1- \mid \kappa \mid^2)^{1/2}_{ap}$. The non local integro-differential equation (2) is changed, as we shall see below, into a system of PDE's. Note that this approximation is designed in such a way that it is valid as long as $c|k|/\omega$ remains small enough, i.e. as long as the wave propagates close enough to the z-direction. We denote by $e(\kappa_1, \kappa_2)$ the corresponding error

$$e(\kappa_1, \kappa_2) = (1- \mid \kappa \mid^2)^{1/2} - (1- \mid \kappa \mid^2)^{1/2}_{ap}.$$

The accuracy of the paraxial approximation depends on the order of magnitude of this error. The so-called fifteen degree paraxial equation, based on the Taylor expansion $\sqrt{1-X} = 1 - X/2 + O(X^2)$, corresponds to an error $e(\kappa_1, \kappa_2) = O(|\ \kappa\ |^4)$ and the forty-five degree paraxial approximation, based on the first Padé expansion $\sqrt{1-X} = \dfrac{1-\frac{3}{4}X}{1-\frac{1}{4}X} + O(X^3)$, to an error $e(\kappa_1, \kappa_2) = O(|\ \kappa\ |^6)$. Figure 1 shows the accuracy of these two classical approximations representing the variations of the error $e(\kappa_1, \kappa_2)$. Typically, the wider is the white area, which represents the zone of directions of propagation where the error is less than 10^{-3}, the better is the approximation.



(a) 15° approximation (parabolic) (b) 45° approximation

Figure 1: **Errors for the classical approximations**

A general class of well-posed high order approximations has been proposed by Lindmann [30] and studied by Bamberger and al [3]. It leads to the following system of (L+1) coupled PDE's , written in the frequency domain (since paraxial equations are usually handled in the frequency domain)

$$
(4) \qquad \left\{ \begin{array}{l} \dfrac{\partial v}{\partial z} + \dfrac{i\omega}{c} v - \dfrac{i\omega}{c} \displaystyle\sum_{\ell=1}^{L} b^\ell \varphi_\ell = 0 \\ \dfrac{\omega^2}{c^2} \varphi_\ell + a^\ell \Delta \varphi_\ell = -\Delta v \qquad \ell = 1, \dots, L. \end{array} \right.
$$

In (4), we still denote by v the solution in the frequency domain. Here L specifies the degree of the approximation, a^ℓ, b^ℓ are positive constants chosen in order to get the best possible approximation and φ_ℓ are auxiliary unknown functions introduced in order to avoid a higher-order PDE.

Classically, one handles the transport term $\dfrac{i\omega}{c} v$ exactly, with the Claerbout change of unknown functions $u = ve^{i\omega z/c}$, and after elimination of the auxiliary unknown functions, end up with the following system (written formally in operator form)

$$
(5) \qquad \left\{ \begin{array}{l} \dfrac{\partial u}{\partial z} - i\dfrac{\omega}{c} \displaystyle\sum_{\ell=1}^{L} A^\ell u = 0 \\ A^\ell = -b^\ell (a^\ell \Delta + \dfrac{\omega^2}{c^2})^{-1} \Delta. \end{array} \right.
$$

Remark 2.1 Actually, to rewrite (4) in the form (5), we should give a precise meaning to the operator $(a^\ell \Delta + \dfrac{\omega^2}{c^2})^{-1}$. This can be done with the help of the limiting absorption principle (see for instance [17]) which is equivalent to use appropriate conditions at infinity.

2.2 Splitting Methods

We briefly recall some classical points about splitting methods (see [33]) which are specifically designed for the solution of ODE's in the form

$$(6) \qquad \left\{ \begin{array}{ll} \dfrac{\partial u}{\partial z}(\boldsymbol{x}, z) - i \displaystyle\sum_{j=1}^{N_s} A_j(z) u(\boldsymbol{x}, z) = 0, \quad z \geq 0, & \boldsymbol{x} = (x_1, x_2) \in \mathbb{R}^2 \\ u(z=0) = u_0 & \text{in } \mathbb{R}^2 \end{array} \right. ,$$

The exact solution of (6) satisfies, forall $z_0 \geq 0$

$$u(z_0 + \Delta z) = \exp(\int_{z_0}^{z_0+\Delta z} i \sum_{j=1}^{N_s} A_j(z) dz) \; u(z_0) \; ,$$

where we use the abusive notation $u(z)$ for denoting the function $\boldsymbol{x} \to u(\boldsymbol{x}, z)$. The splitting methods consist in approximating the exponential of the sum of operators with the product of exponentials

$$(7) \qquad u_{ap}(z_0 + \Delta z) \approx \exp\left(\int_{z_0}^{z_0+\Delta z} i A_{N_s}(z) dz \right) \times \cdots \times \exp\left(\int_{z_0}^{z_0+\Delta z} i A_1(z) dz \right) u(z_0),$$

which is exact when the operators A_j commute (in a homogeneous medium for instance) and second order accurate with respect to Δz when they do not (in laterally inhomogeneous media for instance). Knowing $u(z_0)$, the computation of $u_{ap}(z_0 + \Delta z)$ from expression (7) leads naturally to set $w^0 = u(z_0)$ and to define N_s intermediate unknowns w^j, $j = 1, \ldots, N_s$ satisfying

$$(8) \qquad \left\{ \begin{array}{ll} \dfrac{\partial w_j}{\partial s}(\boldsymbol{x}, s) - i A_j(z_0 + s) w_j(\boldsymbol{x}, s) = 0, \quad 0 \leq s \leq \Delta z, & \boldsymbol{x} \in \mathbb{R}^2 \\ w_j(s=0) = w^{j-1} & \text{in } \mathbb{R}^2 \\ w^j = w_j(\Delta z) & \text{in } \mathbb{R}^2 \end{array} \right. ,$$

w^j being nothing but the result of the product of the j first exponentials applied to $u(z_0)$. In particular we have $u_{ap}(z_0 + \Delta z) = w^{N_s}$. Problem (8), to be solved at each step, is still an evolution problem in depth, but with a single operator.

2.3 The classical numerical schemes

Let us come back to equations (5), using a discrete version of the Laplacian Δ_h on a uniform grid $N \times N$, it would become

$$(9) \qquad \left\{ \begin{array}{l} \dfrac{\partial u}{\partial z} - i \dfrac{\omega}{c} \displaystyle\sum_{\ell=1}^{L} A_h^\ell u = 0 \\ A_h^\ell = -b^\ell (a^\ell \Delta_h + \dfrac{\omega^2}{c^2})^{-1} \Delta_h. \end{array} \right.$$

Remark 2.2 Assume now that the computational domain is bounded and that the solution satisfies for instance Dirichlet boundary conditions on the boundary, thus Δ_h takes into account these boundary conditions. We still have to give a precise meaning to the operator $(a^\ell \Delta_h + \dfrac{\omega^2}{c^2})^{-1}$, forall ℓ, which this time is true if forall ℓ, $\dfrac{\omega^2}{c^2 a^\ell}$ does not coincide with any eigenvalue of the discrete Laplace operator Δ_h.

Since the evolution operator is written as a sum of simpler operators, it is now straightforward to apply splitting methods to this problem. Using for instance an implicit and second accurate Crank-Nicolson scheme for the discretization in depth, we then have to solve a sequence of problems of the following form

$$(10) \qquad \left\{ \begin{array}{l} (\mathrm{I} + d(\omega)\Delta_h)\, u^{k+1} = \left(\mathrm{I} + \overline{d(\omega)}\Delta_h \right) u^k \\ d(\omega) = \dfrac{\alpha c^2}{\omega^2} + i \dfrac{\beta c \Delta z}{2\omega} \end{array} \right. ,$$

which is a linear system, with a large, sparse, complex valued, non-hermitian matrix. Kern [27], [25] has proposed to use modern iterative methods to solve it. We investigate in the following an alternative way to avoid the difficulty altogether, by introducing a new class of paraxial equations suitable for use with splitting in the lateral variables, and requiring only the solution of 1D PDE's. The main outlines of these two alternative ways to solve 3D paraxial equations can be found in [6].

3 The new paraxial approximations

We now explain the technique introduced in [14] for applying splitting methods without loss of accuracy in the approximation. This is done in section 3.1 in the homogeneous whole space. The extension to the heterogeneous case follows in section 3.2 from the criteria given by Bamberger and al. [4]. In view of bounding the computational domain with appropriate absorbing boundaries we present in section 3.3.1 a new technique of designing Perfectly Matched Layers, adapted by Collino to the paraxial wave equations [13] and finally present the 2D model paraxial equation which will be discretized in the next section.

3.1 The homogeneous whole space

Using splitting in the horizontal variables is not a new idea. In order to avoid the inversion of the large matrix when using the full paraxial approximation, Brown [8] has suggested approximating the square root with

$$(1-|\kappa|^2)^{\frac{1}{2}} \approx 1 - \frac{1}{2}\frac{\kappa_1^2}{1-\frac{1}{4}\kappa_1^2} - \frac{1}{2}\frac{\kappa_2^2}{1-\frac{1}{4}\kappa_2^2} \quad (\text{ error : } O(\kappa_1^2\kappa_2^2)),$$

and this has been used by several authors [19, 29, 18]. Unfortunately, this is consistent with the forty-five degree equation (i.e., with an error $\approx O(|\kappa|^6)$) only in the $\kappa_1 = 0$ and $\kappa_2 = 0$ directions. In the other directions, the approximation is of the same order as the 15 degree approximation. Figure 2 shows the error induced by the

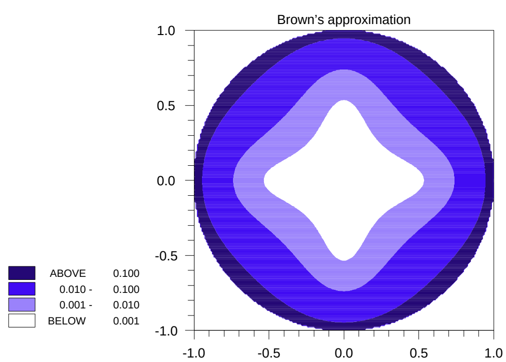


Figure 2: **Error for Brown's approximation**

corresponding approximation. One will note the anisotropy of the result and the loss of accuracy due to the loss of one order in the approximation for directions which are not parallel to the axes.

The basic idea to get better accurate approximations involving, as Brown's approximation, only one lateral space variable per fraction is to introduce more than two directions of splitting. The paraxial equations constructed in Collino and Joly [14] are derived from an approximation of the square root with rational fractions of the following form

$$(1-|\kappa|^2)^{1/2} = (1-(\kappa_1^2+\kappa_2^2))^{1/2} \approx 1 - R(\kappa_1,\kappa_2),$$

with

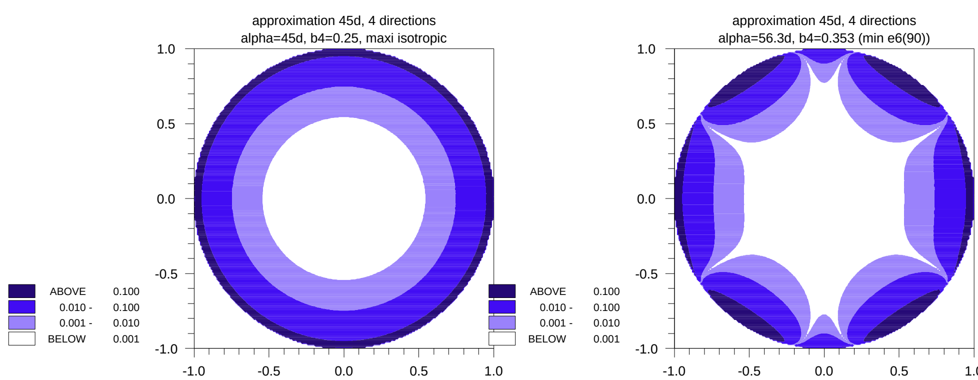
$$R(\kappa) = \sum_{j=1}^{N_D}\sum_{\ell=1}^{L} \frac{b_j^\ell(\kappa.n_j)^2}{1-a_j^\ell(\kappa.n_j)^2}, \tag{11}$$

where N_D corresponds to the number of directions, L the number of fractions per direction, and n_j the unit vector associated to the j^{th} direction ($n_j = (\cos\alpha_j, \sin\alpha_j)$). It has been shown in [14] that the conditions on the coefficients $a_j^\ell > 0$, $b_j^\ell \geq 0$ ensures the well-posedness of these paraxial equations. Also, in order to prevent the approximation error from blowing-up in certain directions inside the unit disk $\kappa_1^2+\kappa_2^2 \leq 1$, it is natural to impose $0 < a_j^\ell \leq 1$. Looking for a given order of accuracy gives rise, via Taylor expansions, to a system that has to be satisfied by the coefficients a_j^ℓ and b_j^ℓ. Several families of approximations of the above type have been constructed so as to achieve comparable accuracy to the classical forty-five ($e(\kappa) = O(|\kappa|^6)$) or sixty degree ($e(\kappa) = O(|\kappa|^8)$) approximations.

Let us present a particular choice: the family of forty-five degree approximations obtained using four directions of splitting, $N_D = 4$, uniformly distributed -namely $x_1, x_1 + x_2, x_1 - x_2, x_2$- and one fraction per direction, $L = 1$. Since there is only one fraction per direction, we omit the superscript and denote by $\{a_j, b_j, j = 1, 2, 3, 4\}$ the coefficients of the approximation. The system satisfied by the coefficients leads to a family of forty-five degree approximations depending on one parameter, i.e. all the coefficients can be expressed in terms of the degree of freedom b_1 that has to be chosen in the interval $[1/12, 5/12]$ in order to ensure the conditions $0 < a_j \leq 1$ and $b_j \geq 0$:

$$\left\{ \begin{array}{l} b_4 = b_1 \;\; ; \;\; b_2 = b_3 = \dfrac{1 - 2b_1}{2} \\ \\ a_1 = a_4 = \dfrac{1}{12 b_1} \;\; ; \;\; a_2 = a_3 = \dfrac{1}{12 b_2} \end{array} \right. , \tag{12}$$

In [14], several criteria are proposed for the choice of the degree of freedom b_1 all based on the error $e(\kappa_1, \kappa_2)$. For instance, the choice $a_j = 1/3, b_j = 1/4, i = 1, \ldots, 4$ gives the "maxi-isotropic" forty-five degree approximation and Figure 3-(a) represents the corresponding error. The quality of the approximation is comparable to the Brown's approximation in the directions κ_1 and κ_2 but the difference is that it remains of the same order in the other directions. More examples are given in the above mentioned paper. In section 6.2, we will propose another criterion for the choice of b_1 (for a squared mesh) based on the dispersion analysis. Note that this particular choice corresponds to the family which is the more adapted to a practical implementation. Indeed, it can be solved numerically on a finite difference mesh (x_1, x_2) of squares of sides $\Delta x_1 = \Delta x_2$ (see below).



(a) **Square mesh ($\Delta x_2 = \Delta x_1$). Error for the maxi isotropic approximation**

(b) **Anisotropic mesh ($\Delta x_2 = 1.5\Delta x_1$) ; Error $e(\kappa_1, \kappa_2)$ for a 45° approximation with b_1 minimizing $|e|$ in the direction κ_2.**

Figure 3: **Errors for the new approximations**

The paraxial equation corresponding to (11) can now be written in a form analogous to (5)

$$\left\{ \begin{array}{l} \dfrac{\partial u}{\partial z} - i\dfrac{\omega}{c} \displaystyle\sum_{j=1}^{N_D} \sum_{\ell=1}^{L} A_j^\ell u = 0 \\ A_j^\ell u = -b_j^\ell (\dfrac{\omega^2}{c^2} + a_j^\ell D_j^2)^{-1} D_j^2 u, \end{array} \right. \tag{13}$$

where $D_j = n_j \cdot \vec{\nabla}$ is the derivative in the j^{th} direction and u is here again linked to the seismic field v via the Claerbout change of unknown $u = v e^{i\omega z/c}$. Each of the operators A_j^ℓ in (13) only involves a one-dimensional differential operator. Thus, this new family of equations lends itself to a splitting method in the horizontal variables, and as mentioned before, this time the splitting has the advantage of being consistent at least with the forty-five degree equation (for a proper choice of coefficients). Denoting again by w^m the intermediate unknowns (see section 2.2), the equations to be solved at each iteration, to go from $w^0 = u(z_0)$ to $w^{N_s} = u(z_0 + \Delta z)$ are the following

$$
(14) \qquad \left\{ \begin{array}{lll} \dfrac{\partial w_m}{\partial s}(\boldsymbol{x}, s) - i\dfrac{\omega}{c}A_m(z_0 + s)w_m(\boldsymbol{x}, s) = 0, & 0 \le s \le \Delta z, & \boldsymbol{x} \in \mathbb{R}^2 \\ w_m(s = 0) = w^{m-1} & \text{in } \mathbb{R}^2 & \\ w^m = w_m(\Delta z) & \text{in } \mathbb{R}^2 & \end{array} \right. ,
$$

where A_m is equal to one of the operators A_j^ℓ.

Remark 3.1 Solving (14) in $\mathbb{R}^2 \times [0, \Delta z]$ is consequently equivalent to solve a continuum of problems posed on lines parallel to the direction n_j. For the numerical exploitation of these equations, for a given direction j, the solution will be computed on a grid $\mathcal{G}^j$ composed on lines parallel to this direction and on nodes located on these lines (see figure 4), so that $D_j u(x_i^j)$ can be approximated with values of u on the line.

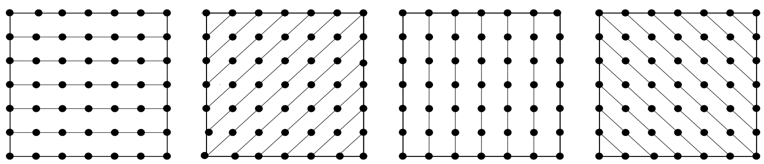


Figure 4: Example of a grid for 4 directions of splitting

Since the solution for the direction j, computed at nodes $x_i^j \in \mathcal{G}^j$ needs the value of the solution for the previous direction $j-1$ at the same nodes, we have the following alternative: either the two grids coincides, so that solution in direction $j-1$ has already been computed on the nodes x_i^j, either they do not and in this case one has to resort to interpolation procedures.

In practice, it is more convenient to work with the same grid for all directions of splitting. This is possible for essentially two choices : (i) take $N_D = 4$ directions with a mesh built from squares or rectangles,(the directions are given by the two coordinate axes and the two main diagonals) ; (ii) take $N_D = 3$ directions with a mesh built from triangles (for equilateral triangles, the 3 directions are 60° apart).

Using more than 4 directions permits to get higher accuracy but implies additional difficulties for the discretization that we do not consider in this paper. Note that 4 directions of splitting are sufficient to get sixty degree approximations provided that there are more than one fraction per direction. Examples are given in [14]. For instance what they call the “cheap equations” necessitate the use of two fractions in two of the four directions and one fraction in the two remaining directions.

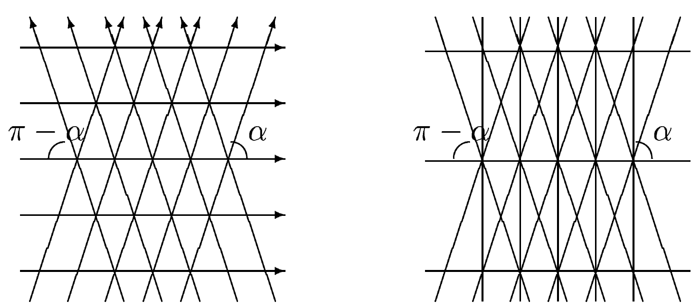


Figure 5: **Left** : anisotropic mesh with 3 directions (the axis x_1 and the two diagonals characterized by the angle α) - **Right** : anisotropic mesh with 4 directions (the two coordinates axes x_1, x_2 and the two diagonals characterized by the angle α)

If we consider the application to geophysical prospecting, the receivers, recording the data at the surface, are localized on a mesh and it may be useful to make this mesh coincide with the one used for the numerical solution of the paraxial equations. Most often, this receivers mesh is what we call an anisotropic mesh, i.e, for which $\Delta x_1 \neq \Delta x_2$. This case has been investigated in [5]. The spatial mesh is composed with non-squared rectangular elements, when using $N_D = 4$ directions (resp. non equilateral triangles, when using $N_D = 3$ directions) and the directions are given, in the case $N_D = 4$, by the two coordinate axes and by the diagonals of the spatial mesh, characterized by the angle α defined as $\Delta x_2/\Delta x_1 = \tan\alpha$ (resp. by only one coordinate axis and by the two diagonals, in the case $N_D = 3$), see figure 5. For such anisotropic meshes analogous families of forty-five degree approximations using 4 directions and 1 fraction per direction have been constructed in [5] provided that the mesh is not "strongly" anisotropic, more precisely under the condition $\alpha \in [\frac{\pi}{6}, \frac{\pi}{3}]$. Again several criteria based on the error $e(\kappa_1, \kappa_2)$ can be proposed in order to choose the degree of freedom. As an illustration, we represent in Figure 3-(b) the error for a mesh such that $\Delta x_2 = 1.5\Delta x_1$, with the degree of freedom chosen in order to minimize the error in the direction κ_2.

3.2 Extension to heterogeneous media

The extension to heterogeneous media is simply obtained by using slight modifications on the operators, following [4]. Actually, paraxial equations in heterogeneous media have been proposed and analyzed by Bamberger and al. [4]. Their approach was to define several criteria (both of mathematical and physical nature), and to select among a general class of possible candidates the one that satisfied those criteria.

Their result gives a recipe which allows one to extend any paraxial equation to heterogeneous media. Thus equations (13) keep the same form, with a new definition for the operators A_j^ℓ

$$\left\{\begin{array}{l} \dfrac{\partial u}{\partial z} - i\dfrac{\omega}{c}\displaystyle\sum_{j=1}^{N_D}\sum_{\ell=1}^{L} A_j^\ell u = 0 \\ A_j^\ell u = -b_j^\ell(\omega^2 + a_j^\ell \Delta_j^c)^{-1}\Delta_j^c u \end{array}\right. , \tag{15}$$

where $\Delta_j^c = cD_j(cD_j)$and c now depends on the position (x_1, x_2, z).

3.3 The equations in a bounded domain

For the numerical computation, the problem has to be set in a cylindrical domain $\mathcal{D} \times \{z \geq 0\}$ bounded in the lateral directions. Either $\mathcal{D}$ is physically bounded with some boundary conditions, as, for instance, Dirichlet or Neumann boundary conditions, or it is unbounded, for instance $\mathcal{D} = \mathbb{R}^2$, and one has to bound it artificially. In the second case, in order to simulate an unbounded domain and to minimize the reflexion of the waves on the boundary, it is crucial to have at one's disposal efficient absorbing boundary conditions. In section 3.3.1 we briefly explain how the Perfectly Matched Layers introduced by Bérenger [7] for Maxwell's equations have been extended to paraxial equations by Collino [13]. In section 3.3.2, we come back to the first case, where $\mathcal{D}$ is bounded with Dirichlet boundary conditions and write down its variational formulation, in view of its discretization with Galerkin type schemes.

3.3.1 Design of absorbing boundary conditions: the PML approach

A problem of practical importance is the treatment of the lateral boundaries, i.e. orthogonal to the z-direction. It must be designed in such a way that the waves are absorbed when they reach the boundaries. Recently, Collino [13] proposed to adapt a new technique, introduced for Maxwell's equations by Bérenger [7]. This technique consists in designing an absorbing layer model called perfectly matched layer (PML). It possesses the astonishing property to generate no reflexion at the interface between the free media and the artificial lossy medium, and the reflected waves are only due to the discretization of the model. This property allows one to use a very high damping parameter inside the layer, and consequently a short layer length, while still achieving a quasi-perfect absorption of the waves.

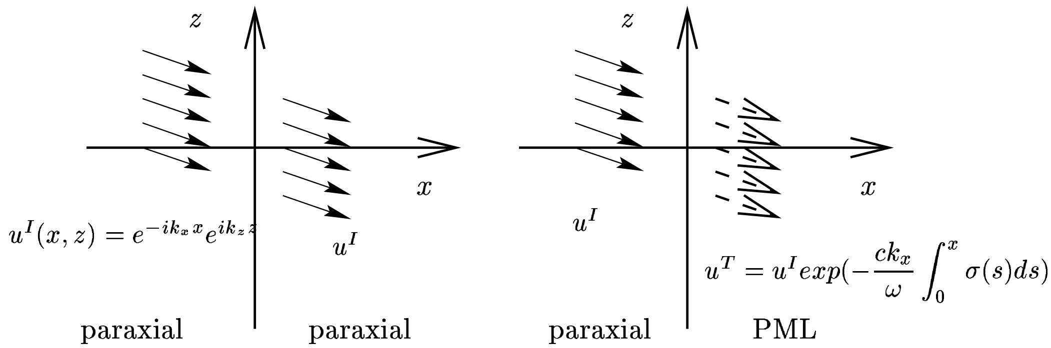


Figure 6: Plane wave propagation. Left: the paraxial model. Right: the paraxial coupled to the PMLs. The plane wave is totally transmitted in the layer and is damped inside the layer.

Consider for instance the simplest 2D paraxial equation, say the fifteen degree approximation

$$\frac{\partial u}{\partial z} + \frac{1}{2}\frac{i}{c\omega}\Delta^c u = 0, \quad \text{in } \mathbb{R}^2 \quad \text{with } \Delta^c = c\partial_x c\partial_x \tag{16}$$

For designing our PMLs model in the right half-space $x > 0$, we define a positive function $\sigma(x)$ with support in the damped area (i.e., $\sigma(x) = 0$ for $x < 0$ and $\sigma(x) = |\sigma(x)|$ for $x > 0$) and a new model

$$\frac{\partial \widetilde{u}}{\partial z} + \frac{1}{2}\frac{i}{c\omega}\widetilde{\Delta}^c \widetilde{u} = 0, \quad \text{in } \mathbb{R}^2 \quad \text{with } \widetilde{\Delta}^c = cd(x)\partial_x cd(x)\partial_x \tag{17}$$

and $d(x) = \dfrac{i\omega}{i\omega + c\sigma(x)}$. The condition $\sigma(x) \equiv 0$ for $x < 0$ implies that the two waves u and $\widetilde{u}$ coincide in the left half-space, $\widetilde{u}(x,z) = u(x,z)$, $\forall x < 0$, they differ only in the damped half-space where $\widetilde{u}$ is exponentially damped. This property can be seen through a plane wave analysis. Consider the incident right-going plane wave $u^I(x,z) = e^{-ik_x x}e^{ik_z z}$ with $k_x > 0$ for model (16), (i.e., satisfying the dispersion relation $k_x^2 = 2\frac{\omega}{c}k_z$), see Fig. 6. It is also solution of (17) in the left half-space. Taking this wave as an incident wave in (17) it can be shown that there is no reflection, the wave is totally transmitted in the layer, and the transmitted wave is really damped inside the layer, i.e.,

$$\widetilde{u} = \begin{cases} u^I + u^R & \text{in } x < 0 \\ u^T & \text{in } x > 0 \end{cases} \Longrightarrow \begin{cases} u^R = 0 \\ \left|u^T(x,z)\right| = \left|u^I(x,z)\right| exp\left(-\dfrac{ck_x}{\omega}\displaystyle\int_0^x \sigma(s)ds\right) \end{cases}$$

For the practical implementation, we cannot work with the absorbing half-space $x > 0$. We have to deal with a finite lenght absorbing layer. If δ is the length of the layer, the system of equations is closed with a Dirichlet boundary condition which produces a reflected wave. The same plane wave analysis shows that the reflection coefficient is given by

$$R = exp\left(-2\frac{ck_x}{\omega}\int_0^\delta \sigma(s)ds\right)$$

which describes the percentage of the original wave amplitude after two passes through the PML.

Practically, the PML is very easy to implement with the 2D paraxial equation since it consists only to change the operator Δ^c to the operator $\widetilde{\Delta}^c$. Concerning the application to the new 3D paraxial equations, we have to solve a series of 2D paraxial equations in each direction, and it is natural to use the same process applied to each 2D problem (changing Δ_j^c to $\widetilde{\Delta}_j^c$). However, it is not so clear to interprete exactly what problem is solved in the global layer around the domain of interest, Ω_i, in the (x,y)-plane (see figure 7). If we consider for instance the 45 degree approximation using 4 directions of splitting, and apply the technique described for the 2D paraxial equations, on each line on which we have a 2D problem to solve, the point is to know how to define $\sigma(x_i^j)$ on each line, in order to change Δ_j^c to $\widetilde{\Delta}_j^c$. First denote by $\Sigma(s)$ the profile for the damping in a 1D layer of width δ, located in $[0,\delta]$ (see Fig. 7). We proceed as follows (see figure 7):

• For the x-direction (resp. y), we do as if there was absorbing layers only in the orthogonal direction, and on each line we put the profile Σ (at the extremities).

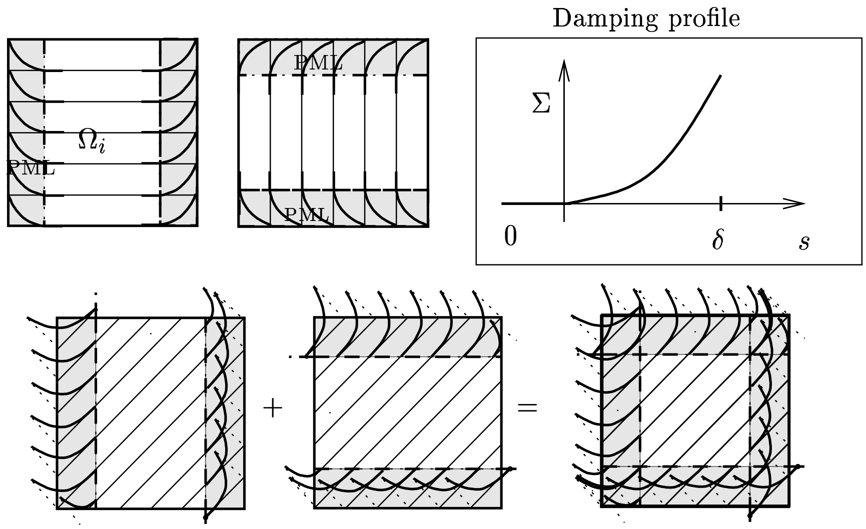


Figure 7: Description of the damping on each line

• For the diagonal directions, $x-y$ for instance, we do as if there was absorbing layers successively in x and in y and finally add both contributions.

In section 7, numerical experiments obtained with these PMLs and with a Dirichlet boundary condition are compared.

3.3.2 The 2D model paraxial equation with Dirichlet boundary conditions

In this section, we write down in more details the serie of 2D paraxial equations we end up after using the splitting method. For the sake of simplicity, (i) we deal with the approximations using only one fraction per direction ($L=1$), the extension to several fraction being straightforward, (ii) we consider these problems with Dirichlet boundary conditions on the lateral boundaries. All these problems can be analyzed through a model 2D paraxial equation, described in paragraph **b**.

a - Reduction to a series of 2D paraxial equations

After splitting, at each iteration, i.e. to go from z_0 to $z_0+\Delta z_0$, and for each direction $1 \le j \le N_D$, the equation (14) has to be solved in order to determine the j-th intermediate unknown w_j. This can be rewritten in a heterogeneous medium, introducing an auxiliary unknown $\varphi_j(\boldsymbol{x},s) = A_j(\boldsymbol{x}, z_0+s)w_j(\boldsymbol{x},s)$, $0 \le s \le \Delta z$, as

$$
(18) \qquad \left\{
\begin{array}{ll}
\dfrac{\partial w_j}{\partial s}(\boldsymbol{x},s) - i\dfrac{\omega}{c}\varphi_j(\boldsymbol{x},s) = 0, & 0 \le s \le \Delta z, \quad \boldsymbol{x} = (x_1,x_2) \in \mathcal{D} \\[2ex]
\dfrac{\omega^2}{c}\varphi_j + D_j\left(cD_j(a_j\varphi_j + b_j w_j)\right) = 0, & (\boldsymbol{x},s) \in \mathcal{D} \times [0,\Delta z] \\[2ex]
w_j(s=0) = w^{j-1}, & \text{in} \quad \mathcal{D} \\[2ex]
w^j = w_j(\Delta z). &
\end{array}
\right.
$$

Let $\mathcal{G}^j$ be, as in section 3.1, a grid composed on lines parallel to direction j (see figure 4). Problem (18) can be solved on each line independently, knowing the solution coming from the previous direction. Considering one particular line, parameterized by a parameter $\tau \in \Omega =]-L,+L[$, from remark 3.1, we easily see that this can be rewritten as a one-dimensional problem, on the line

$$(19)\qquad \left\{\begin{array}{ll} \dfrac{\partial w_j}{\partial s}(\boldsymbol{x}(\tau),s) - i\dfrac{\omega}{c}\varphi_j(\boldsymbol{x}(\tau),s) = 0, & \text{in } \Omega\times[0,\Delta z] \\ \dfrac{\omega^2}{c}\varphi_j + \dfrac{\partial}{\partial\tau}\left(c\dfrac{\partial}{\partial\tau}(a_j\varphi_j + b_j w_j)\right) = 0, & \text{in } \Omega\times[0,\Delta z] \\ w_j(s=0) = w^{j-1}, & \text{in } \ \Omega \\ w^j = w_j(s=\Delta z), & \text{in } \ \Omega. \end{array}\right.$$

Problems (19) are nothing but 2D paraxial equations, w_j being related as usual to the 2D seismic field v_j via the Claerbout change of unknown.

b - The 2D model paraxial equation

Without loss of generality, we focus on a particular direction, that we denote by x. The domain Ω is again an interval in $\mathbb{R}$, $\Omega =]-L,+L[$, and we want to solve the following equations (where $a>0$ and $b\geq 0$)

$$(20)\qquad \left\{\begin{array}{lll} \dfrac{\partial w}{\partial z} - \dfrac{i\omega}{c}\varphi = 0, & \text{in } \Omega\times[0,Z] & (a) \\ \dfrac{\omega^2}{c}\varphi + \dfrac{\partial}{\partial x}\left(c\dfrac{\partial}{\partial x}(a\varphi + bw)\right) = 0, & \text{in } \Omega\times[0,Z] & (b) \\ w(z=0) = w_0, & \text{in } \Omega & (c), \end{array}\right.$$

and w is as usual related to the 2D seismic wave field v via the Claerbout change of unknown

$$(21)\qquad w = e^{i\frac{\omega}{c}z}v.$$

Note that the space derivatives only concerns the function $\psi \equiv a\varphi + bw$. It is therefore natural to ask more regularity in the space variable on ψ compared to w and φ. To close this system, we need for ψ an additional boundary condition on the lateral boundaries. We consider here the Dirichlet boundary condition

$$(22)\qquad \psi = 0 \quad \text{on } \partial\Omega\times[0,Z].$$

To give a variational formulation of (20)-(22), we introduce the following functional framework (see for instance [31], for the definitions of Sobolev spaces) $H = L^2(\Omega)$, $V = H_0^1(\Omega)$. We denote the inner product and the norm in H by

$$(u,v)_{0,\Omega} = \int_\Omega u\bar{v}\,dx, \quad \|u\|_{0,\Omega} = (u,u)_{0,\Omega}^{1/2},$$

the semi-norm in V by

$$|\phi|_{1,\Omega} = \left(\int_\Omega \left|\frac{\partial\phi}{\partial x}\right|^2 dx\right)^{1/2},$$

which is equivalent on V to the usual norm $\|\,\phi\,\|_{1,\Omega} = \left(\|\phi\|_0^2 + |\phi|_{1,\Omega}^2\right)^{1/2}$. We set

$$(23)\qquad m(u,v) = \int_\Omega \frac{1}{c}u\bar{v}\,dx, \quad (u,v)\in H\ ;\ k(u,v) = \int_\Omega c\frac{\partial u}{\partial x}\frac{\partial\bar{v}}{\partial x}\,dx, \quad (u,v)\in V\ ,$$

$m(.,.)$ is called the mass bilinear form and $k(.,.)$ the stiffness bilinear form. In the following the velocity $c(x,z)$ is assumed to be regular enough and we denote by c_m and c_M the lower and upper bounds

$$c_m \leq c(x,z) \leq c_M \quad \forall (x,z)\in\overline{\Omega}\times[0,Z]\ .$$

This leads obviously to the following bounds

$$(24)\qquad \left\{\begin{array}{l} \dfrac{1}{c_M}\,\|\,\psi\,\|_{0,\Omega}^2 \leq m(\psi,\psi) \leq \dfrac{1}{c_m}\,\|\,\psi\,\|_{0,\Omega}^2 \\ c_m|\psi|_{1,\Omega}^2 \leq k(\psi,\psi) \leq c_M|\psi|_{1,\Omega}^2 \end{array}\right. ,$$

which shows that (i) $m(.,.)$ induces a norm on L^2, (ii) $k(.,.)$ induces a norm on H_0^1.

We assume that $w_0 \in H$. With these definitions, problem (20) admits the variational formulation : find $(w,\varphi,\psi) \in C^0(0,Z;H) \times L^2(0,Z;H) \times L^2(0,Z;V)$ such that

$$(25) \quad \left\{ \begin{array}{ll} \dfrac{d}{dz}(w,\chi) - i\omega m(\varphi,\chi) = 0, & \forall\chi \in H \\ \omega^2 m(\varphi,\phi) - k(\psi,\phi) = 0, & \forall\phi \in V \\ \psi = a\varphi + bw & \text{in } V \\ w(z=0) = w_0. & \end{array} \right. ,$$

where the derivative $\dfrac{d}{dz}$ is taken in the distributional sense. We need w to be $C^0(0,Z;H)$ and not only $L^2(0,Z;H)$ in order to give a meaning to the initial condition $w(z=0)$.

We recall an important and classical property of the solution to (25) (see for instance Kern [24])

Proposition 3.1 *Any solution $(w,\varphi,\psi) \in C^0(0,Z;H) \times L^2(0,Z;H) \times L^2(0,Z;V)$ to (25) satisfies the conservation of energy*

$$(26) \quad \| w(z) \|_{0,\Omega} = \| w_0 \|_{0,\Omega}, \quad \forall z \,.$$

This result is of course important since it shows the L^2-continuity of the solution w with respect to the data w_0 and uniqueness of the solution (w,φ,ψ) follows

Proposition 3.2 *If problem (25) admits a solution $(w,\varphi,\psi) \in C^0(0,Z;H) \times L^2(0,Z;H) \times L^2(0,Z;V)$, this solution is unique.*

Finally, when the frequency does not coincide with an irregular frequency, i.e. a frequency for which there exists $\Psi \neq 0 \in V$ such that

$$\frac{\omega^2}{a} m(\Psi,\phi) - k(\Psi,\phi) = 0 \quad \forall\phi \in V,$$

we have the existence result

Proposition 3.3 *If ω does not coincide with an irregular frequency then problem (25) admits a solution $(w,\varphi,\psi) \in C^0(0,Z;H) \times L^2(0,Z;H) \times L^2(0,Z;V)$.*

4 Higher-order schemes for a 2D paraxial equation

4.1 Semi-discretization in depth

Problem (20) is an evolution problem in depth. We assume here the velocity to be independent on z between two consecutive steps z^m and z^{m+1} ($c(x_1,x_2,z) = c^m(x_1,x_2)$ for $z \in [z^m, z^{m+1}]$), and rewrite (20) in the interval $[z^m, z^{m+1}]$

$$(27) \quad \left\{ \begin{array}{l} \dfrac{dw}{dz} = iC^m w \quad z^m \le z \le z^{m+1} \\ w(z^m) = w^m \end{array} \right. ,$$

with

$$(28) \quad C^m = -\frac{\omega}{c^m}(\omega^2 + a\Delta_m^c)^{-1} b \Delta_m^c \,,$$

and $\Delta_m^c = c^m \dfrac{\partial}{\partial x}\left(c^m \dfrac{\partial}{\partial x}\right)$. We assume again that we have been able to eliminate the auxiliary unknown which means here that $\dfrac{\omega^2}{a}$ does not coincide with an eigenvalue of the operator Δ_m^c with Dirichlet boundary conditions. We use the discretization in depth initially proposed by Joly and Kern [24] (1990). It is based on the expression of the exact solution of (27), $w(z^{m+1}) = e^{iC^m \Delta z} w(z^m)$, and on the relationship between Runge-Kutta methods and Padé approximations to the exponential (see for instance Hairer and Wanner (1991)[20]). In order to get conservative schemes of order $2K$, the exponential is replaced by a Padé approximant on the following form

$$\prod_{k=1}^{K} \frac{1+r_k x}{1+\bar{r}_k x} = \frac{N_K(x)}{\overline{N}_K(x)} \ , \tag{29}$$

where the coefficients r_k are chosen in such a way that

$$\exp(ix) = \frac{N_K(x)}{\overline{N}_K(x)} + O(|x|^{2K}) \ .$$

The integration from z^m to z^{m+1} is then formally done as follows

$$w^{m+1} = \prod_{k=1}^{K} (I + \bar{r}_k \Delta z C^m)^{-1} (I + r_k \Delta z C^m) w^m \ . \tag{30}$$

This procedure leads us, as in the splitting process, to define K intermediate unknowns, denoted by w_k^m, associated to each fraction and defined by

$$w_0^m = w^m, \quad (I + \bar{r}_k \Delta z C^m) w_k^m = (I + r_k \Delta z C^m) w_{k-1}^m, \quad 1 \le k \le K.$$

We then set $w^{m+1} = w_K^m$.

The classical Crank-Nicolson second-order scheme is obtained for $K = 1$ and $r_1 = i/2$. Of course, the addition of intermediary steps increases the cost of these schemes : the cost of the 2K order scheme is about K times the cost of the 2nd-order scheme, as there are K systems to solve instead of one.

Remark 4.1 In Collino's thesis [12], it was found that use of the θ-scheme, with $\theta > 1/2$ could be beneficial to suppress unwanted components of the wavefront. For this reason, Kern [26](1992) has shown that this process can also be used to get non conservative schemes, which generalize the θ-scheme to higher order schemes.

4.2 Discretization in the lateral variable with higher-order variational finite difference schemes

This section is devoted to the construction of higher-order semi-discretization in x of system (20). We base the discretization on the variational formulation (25), which provides us with a systematic treatment of heterogeneous media, in a way that insures good numerical properties (stability thanks to energy estimates). The most common way to approximate (25) is to use finite elements P_k (see for instance [2, 1]). This would yield several drawbacks. First of all, although this section is concerned with the 2D model paraxial equation, one has to keep in mind that it is only one step in the use of the splitting method to get the 3D solution. According to remark 3.1, the solution computed in the direction j at points x_i^j needs the value of the solution computed in the previous direction $j-1$ at the same points. For the lowest-order P_1 finite elements there is no difficulty since all the degrees of freedom coincide with the mesh points. This is not true anymore for higher-order P_k finite elements, $k \ge 2$, since there are additional degrees of freedom between two mesh points. These new degrees of freedom do not coincide in each direction, therefore the evaluation of the solution at these points would require the use of interpolation procedures. Moreover, since all the nodes do not play the same role, the construction of the schemes is not so straightforward as for finite difference type schemes.

This is why we have preferred to use a variational finite difference approach. It is still based on the variational approach and thus keeps the advantage of the systematic treatment of heterogeneous media in a stable way. A first family of schemes, the so-called classical schemes, are presented in section 4.2.3. We then extend to higher-order schemes a classical idea used for lowest-order schemes (see Claerbout [11], Collino [12]) in order to get a more accurate scheme, using a slight modification of the classical one. It is essentially based on the use of an approximation of the mass matrix which allows us to gain two orders of accuracy compared with the classical discretization. This leads in section 4.2.4 to the family of so-called modified schemes. Note that with a finite element approach, this technique amounts to using an approximate mass matrix coming from a mass-lumped mass matrix, i.e. an approximation of the exact mass matrix based on the use of quadrature formulas. Mass-lumping is feasible provided that the quadrature points coincide with the degrees of freedom. Moreover, if we want to keep the same order of accuracy, we know from [10] that the use of P_k finite elements requires quadrature formulas exact on polynomials of degree $2k-1$. It has been shown by N. Tordjman [38] that these two requirements are not always possible. For instance with P_3 finite elements, the mass-lumping is not possible with the classical degrees of freedom. This difficulty is another argument in favour of the use of the finite difference variational approach.

4.2.1 Presentation of the discretization

The problem to approximate can be expressed as follows : find $(w, \varphi) \in C^0(0, Z; H) \times L^2(0, Z; H)$ such that $a\varphi + bw \in L^2(0, Z; V)$ and satisfying

$$(31) \qquad \left\{ \begin{array}{ll} \dfrac{d}{dz}(w, \chi) - i\omega m(\varphi, \chi) = 0, & \forall \chi \in H \quad (a) \\ \omega^2 m(\varphi, \phi) - k(a\varphi + bw, \phi) = 0, & \forall \phi \in V \quad (b) \\ w(z = 0) = w_0 & \end{array} \right. .$$

The discretization is built upon a variational finite difference method. The domain is discretized by a regular grid ($x_i = i\Delta x$) and we define the shifted grid by the nodes ($x_{i+1/2} = (i + 1/2)\Delta x = (i + 1/2)h$), see figure 8.

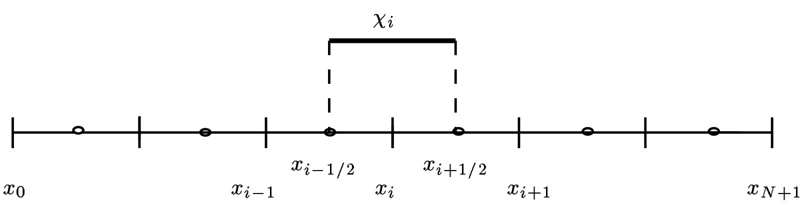


Figure 8: χ_i, basis of H_h

We look for an approximate solution $(w_h, \varphi_h) \in (C^0(0, Z; H_h))^2$ where H_h is the finite dimension space

$$H_h = \left\{ v_h \in L^2(\Omega); v_{h/[x_{i-1/2}, x_{i+1/2}]} \in P^0 \right\},$$

which is only included in $L^2(\Omega)$. A function v_h of H_h is therefore characterized by its values at nodes x_i. For the approximate solution, the constrain $a\varphi_h + bw_h \in L^2(0, Z; V)$ is a priory not satisfied, which means that we use a non-conforming approximation. However, the stiffness bilinear form $k(.,.)$ involving the lateral derivative terms would need more regularity on $a\varphi_h + bw_h$, therefore this term is approximated by a bilinear form $k_h(.,.)$ defined on H_h (see below). Let $(\chi_i)_{i=1,\dots,N}$ be the basis of H_h defined as

$$\chi_i(x_j) = \frac{1}{\sqrt{h}} \delta_{ij} \quad 1 \le i, j \le N \ ,$$

(the factor $1/\sqrt{h}$ is introduced in order to get the normalization $(\chi_i, \chi_j)_0 = \delta_{ij}$), χ_i corresponds to the characteristic function on $[x_{i-1/2}, x_{i+1/2}]$, apart from the factor $1/\sqrt{h}$. For any function v_h of H_h, we let v_i its components in this basis, $(v_h, \chi_i)_0 = v_i = \sqrt{h}\, v_h(x_i)$ and V_h the vector of components v_i. There is a canonical bijection between H_h and $\mathbb{C}^N$. In particular if $\|\ .\ \|$ denotes the norm in $\mathbb{C}^N$ and $\bullet$ the vector scalar product, we have $\| v_h \|_0 = \| V_h \|$ and for $(v_h, t_h) \in H_h^2$ of components $(V_h, T_h) \in (\mathbb{C}^N)^2$: $(v_h, t_h)_0 = V_h \bullet T_h$. For a regular function $\rho \in C^0(\overline{\Omega})$ we define its interpolation $\pi_h \rho$ as the function in H_h such that $\pi_h \rho(x_i) = \rho(x_i)\ \forall i$, or equivalently as

$$(32) \qquad \pi_h \rho = \sqrt{h} \sum_i \rho(x_i) \chi_i \ .$$

The approximate solution is decomposed on the basis as foll ows

$$w_h(x, z) = \sum_{i=1}^{N} W_i(z) \chi_i(x) \quad ; \quad \varphi_h(x, z) = \sum_{i=1}^{N} \Phi_i(z) \chi_i(x) \ .$$

We denote by W_h and Φ_h the unknown vector functions $(W_h)_i = (W_i)_{1 \le i \le N}$ and $(\Phi_h)_i = (\Phi_i)_{1 \le i \le N}$ and we make the convention $W_0 = W_{N+1} = \Phi_0 = \Phi_{N+1} = 0$. The approximate problem then consists in finding the vector functions (W_h, Φ_h) satisfying the matricial system

$$(33) \qquad \left\{ \begin{array}{l} \dfrac{dW_h}{dz} - i\omega M_h \Phi_h = 0 \\ \left(\omega^2 M_h - aK_h\right) \Phi_h = bK_h W_h \end{array} \right. ,$$

where M_h is a diagonal matrix, called the mass matrix, K_h is the stiffness matrix

$$(M_h)_{ij} = m(\chi_i, \chi_j) \quad ; \quad (K_h)_{ij} = k_h(\chi_i, \chi_j) \ ,$$

and k_h will be explicited below. The correspondence between H_h and $\mathbb{C}^N$ yields $M_h V_h \bullet T_h = m(v_h, t_h)$ and $K_h V_h \bullet T_h = k_h(v_h, t_h)$. Properties of the bilinear form $m(.,.)$ implies that M_h is a symmetric (diagonal !) definite positive matrix. The scheme (33) will be completely defined only once k_h is known.

4.2.2 Definition of the approximate stiffness bilinear form k_h

To derive an approximate stiffness bilinear form, the derivative operator $\partial/\partial x$ is approximated with a finite difference operator defined on H_h. Let D_ε denote the usual second-order finite difference approximation of d/dx (apart from the factor $1/\varepsilon$), i.e.,

$$D_\varepsilon \phi(x) \equiv \phi(x+\varepsilon) - \phi(x-\varepsilon),$$

and $D_\varepsilon^\star$ its adjoint, which is simply defined as $D_\varepsilon^\star = -D_\varepsilon$. For $\phi \in H_h$, and $\varepsilon = h/2$, $\phi(x)$ is defined on each interval $[x_{i-1/2}, x_{i+1/2}]$, thus $\phi(x+h/2)$ is defined on the shifted intervals $[x_{i-1}, x_i]$ and we have $D_{h/2}\phi(x) = \phi(x_i) - \phi(x_{i-1})$ for $x \in [x_{i-1}, x_i]$, $\forall 1 \leq i \leq N+1$ (with $\phi(x_0) = \phi(x_{N+1}) = 0$). This shows that $D_{h/2}\phi$ belongs to $H_h^{1/2}$ which is the set of functions that are constants on each interval $[x_{i-1}, x_i]$ or equivalently of functions characterized by their values at nodes of the shifted grid $x_{i-1/2}$

$$H_h^{1/2} = \left\{ v_h \in L^2(\Omega) \text{ such that } v_{h/[x_{i-1},x_i]} \in P^0, \forall 1 \leq i \leq N+1 \right\} \ .$$

More generally, if we consider the finite difference operator $D_{(2p-1)h/2}$, it is also defined on the shifted intervals, i.e., it still belongs to $H_h^{1/2}$ (with the convention $\phi(x_j) = 0$ for $j \leq 0$ and $j \geq N+1$). We define a finite difference operator approximating d/dx as a linear combination on the following form

$$\partial_h = \frac{1}{h} \sum_{p=1}^{n} \nu_p D_{(2p-1)h/2} \ . \tag{34}$$

In particular we have $\partial_h \phi(x_{j+1/2}) = \dfrac{1}{h} \displaystyle\sum_{p=1}^{n} \nu_p \left(\phi(x_{j+p}) - \phi(x_{j-p+1})\right)$ and again for a function $\phi \in H_h$, $\partial_h \phi$ is in $H_h^{1/2}$. Note that to approximate a second order derivative, we would apply again this operator and thus evaluate this second order derivative at the nodes of the original grid. This procedure allows us to recover some finite difference schemes used classically when the medium is homogeneous.

This approximation will be said of *order* r if, for a sufficiently regular function ρ, the error between the approximate and the exact derivative is of order r, i.e., if

$$\partial_h \rho(x) = \frac{d\rho}{dx}(x) + O(h^r), \quad \forall x \ ,$$

and of course in this case $\partial_h^2 \rho$ is an approximation of $\dfrac{d^2\rho}{dx^2}$ of order r.

Lemma 4.1 *The approximation given by (34) is of order* $2n$ *provided that the coefficients* ν_p *satisfy the system*

$$\sum_{p=1}^{n} \nu_p (2p-1)^{2k-1} = \delta_{k1} \quad \textit{for } 1 \leq k \leq n \ . \tag{35}$$

In that case it is denoted by $\partial_h^{[2n]}$ *and for any regular function* ρ *we have*

$$\partial_h^{[2n]} \rho(x) = \frac{d\rho}{dx}(x) + h^{2n} R_S^{[2n]} \rho^{(2n+1)}(x) + O(h^{2n+2}) \ , \tag{36}$$

where

$$R_S^{[2n]} = \frac{1}{(2n+1)!\ 2^{2n}} \sum_{p=1}^{n} \nu_p (2p-1)^{2n+1} \ . \tag{37}$$

Proof : let ρ be in $C^{2n+1}(\overline{\Omega})$, $D_\varepsilon \rho$ admits the following Taylor expansion

$$D_\varepsilon \rho(x) = 2\sum_{p=1}^{n} \frac{\varepsilon^{2p-1}}{(2p-1)!}\rho^{(2p-1)}(x) + O(\varepsilon^{2n+1}) \ ,$$

and this yields

$$\partial_h \rho(x) = \sum_{k=1}^{n} \frac{h^{2k-2}}{(2k-1)!2^{2k-2}}\rho^{(2k-1)}(x)\sum_{p=1}^{n}\nu_p(2p-1)^{2k-1} + O(h^{2n}) \ .$$

The condition (35) on the coefficients to get a $2n$-order approximation follows then easily from this expression. Expliciting the remainder, we get

$$\partial_h^{[2n]}\rho(x) = \frac{d\rho}{dx}(x) + h^{2n}\frac{\rho^{(2n+1)}(x)}{(2n+1)!\ 2^{2n}}\sum_{p=1}^{n}\nu_p(2p-1)^{2n+1} + O(h^{2n+2}) \ . \qquad \blacksquare$$

This leads to the approximate stiffness bilinear form (corresponding to the $2n$-order approximation)

$$k_h^{[2n]}(\phi,\chi) = (c\partial_h^{[2n]}\phi, \partial_h^{[2n]}\chi), \quad \forall(\phi,\chi)\in H_h^2 \ ,$$

which is symmetric and satisfies $c_m \parallel \partial_h^{[2n]}\phi \parallel_0^2 \leq k_h^{[2n]}(\phi,\phi) \leq c_M \parallel \partial_h^{[2n]}\phi \parallel_0^2$, $\forall \phi\in H_h$. The stiffness matrix in the matricial system (33), when using the $2n$-order approximation, is simply defined as

$$(K_h)_{ij}^{[2n]} = k_h^{[2n]}(\chi_i,\chi_j) = (c\partial_h^{[2n]}\chi_i, \partial_h^{[2n]}\chi_j) \ .$$

4.2.3 The classical schemes

a - Description

With the above definitions, we now define the $2n$**-order classical scheme** as follows

$$(38) \qquad \left\{ \begin{array}{ll} \dfrac{dW_h}{dz} - i\omega M_h \Phi_h = 0 & (a) \\ \left(\omega^2 M_h - aK_h^{[2n]}\right)\Phi_h = bK_h^{[2n]}W_h & (b) \end{array} \right. .$$

In practice, the auxiliary unknown is eliminated in order to get an evolution system on W_h (this is possible if we assume that $\dfrac{\omega^2}{a}$ is not an eigenvalue of the matrix $(M_h)^{-1}K_h^{[2n]}$) and it gives

$$(39) \qquad \left(\omega^2 M_h - aK_h^{[2n]}\right)(M_h)^{-1}\frac{dW_h}{dz} = i\omega b K_h^{[2n]}W_h \ .$$

The left hand side matrix necessitates to invert M_h which is easy since M_h is a diagonal matrix. This property is important to keep in mind when constructing the new modified schemes.

b - Order of the classical schemes in a homogeneous medium

As mentioned previously, the scheme (38) can be interpreted in terms of a finite difference scheme. We analyze its order in the case of a homogeneous medium. In this case, we rewrite (38) as follows

$$(40) \qquad \left\{ \begin{array}{ll} \dfrac{dw_h}{dz}(x_j,z) - \dfrac{i\omega}{c}\varphi_h(x_j,z) = 0, \ \ \forall j & (a) \\ \dfrac{\omega^2}{c}\varphi_h(x_j,z) - \displaystyle\sum_i (K_h^{[2n]})_{ji}\left(a\varphi_h(x_i,z) + bw_h(x_i,z)\right) = 0 & (b) \end{array} \right. .$$

To analyze the "quality" of this approximation, we choose as a criterion of quality the truncation error which quantifies at which order the exact solution (w,φ) of (20) satisfies the scheme. The first equation is trivially exactly satisfied by the exact solution. The error comes thus from the second equation and is defined as

$$(41) \qquad E_h^{class} = \frac{\omega^2}{c}\varphi(x_j,z) - \sum_i (K_h^{[2n]})_{ji}(a\varphi(x_i,z) + bw(x_i,z)) \ .$$

Lemma 4.2 *The scheme (38) is of order 2n. The first term of the truncation error can be explicited*

$$E_h^{class} = 2cR_S^{[2n]} h^{2n} \frac{\partial^{2n+2}\psi}{\partial x^{2n+2}}(x_j, z) + O(h^{2n+2}) \ ,$$

where R_S has been defined in (37) and $\psi = a\varphi + bw$.

Proof : see appendix A. ■

4.2.4 The modified schemes

a - Description

The left hand side matrix of the classical scheme (39), $\omega^2 M_h - aK_h^{[2n]}$, has the same bandwidth as $K_h^{[2n]}$. Could we modify this matrix in such a way that the bandwidth remains the same and that we gain some orders of accuracy ? The answer is yes and is given by what we call **the modified schemes**. The idea is to introduce a new mass matrix, that we already denote by $M_\alpha^{[2n]}$ (see below), instead of M_h, which has the same bandwidth as $K_h^{[2n]}$, so that the global bandwidth of $\omega^2 M_\alpha^{[2n]} - aK_h^{[2n]}$ does not change. The new system to solve would then be

(42) $$\left(\omega^2 M_\alpha^{[2n]} - aK_h^{[2n]}\right)(M_h)^{-1}\frac{dW_h}{dz} = i\omega bK_h^{[2n]} W_h \ .$$

This corresponds to modify the mass matrix only in the second equation (b) of (38) and to solve

(43) $$\left\{\begin{array}{ll} \dfrac{dW_h}{dz} - i\omega M_h \Phi_h = 0 & (a) \\ \left(\omega^2 M_\alpha^{[2n]} - aK_h^{[2n]}\right)\Phi_h = bK_h^{[2n]} W_h & (b') \end{array}\right. .$$

As explained below, this process can be seen as an extension of Claerbout's scheme (1983) [11] to higher orders.

b - Construction of $M_\alpha^{[2n]}$

Recall the definition of the mass matrix

$$(M_h)_{ij} = (\frac{1}{c}\chi_i, \chi_j)_0 \ .$$

A natural way to construct an approximation of M_h is to approximate the operator identity I in an analogous way than we have approximated $\partial/\partial x$. We follow the same kind of procedure as for the construction of the stiffness matrix. We set I_ε the finite difference operator approximating the identity

$$I_\varepsilon \phi(x) = \frac{1}{2}(\phi(x+\varepsilon) + \phi(x-\varepsilon)) \ ,$$

which is self adjoint, $I_\varepsilon^\star = I_\varepsilon$. Again for a function $\phi \in H_h$ and for $\varepsilon = (2p-1)h/2$ the resulting function $I_\varepsilon\phi$ belongs to $H_h^{1/2}$. We introduce a higher order approximation of I, δ_h, as a linear combination

(44) $$\delta_h = \sum_{p=1}^{n} \mu_p I_{(2p-1)h/2} \ .$$

Lemma 4.3 *The approximation given by (44) is of order 2n -i.e., $\delta_h\rho(x) = \rho(x) + O(h^{2n})\ \forall x$, for any regular function ρ- provided that the coefficients μ_p satisfy the Vandermonde system*

(45) $$\sum_{p=1}^{n} \mu_p (2p-1)^{2(k-1)} = \delta_{k1} \quad \text{for } 1 \le k \le n \ .$$

In this case, it is denoted $\delta_h^{[2n]}$ and for any regular function ρ we have

(46) $$\delta_h^{[2n]}\rho(x) = \rho(x) + h^{2n} R_M^{[2n]} \rho^{(2n)}(x) + O(h^{2n+2}) \ ,$$

where

(47) $$R_M^{[2n]} = \frac{1}{(2n)!\ 2^{2n}} \sum_{p=1}^{n} \mu_p (2p-1)^{2n} \ .$$

Proof : see appendix B. ■

We are now able to construct an approximate mass matrix

$$(U_h^{[2n]})_{ij} = (\frac{1}{c}\delta_h^{[2n]}\chi_i, \delta_h^{[2n]}\chi_j)_0 \ , \tag{48}$$

which is clearly symmetric and positive ($U_h^{[2n]}V_h \bullet V_h = (\frac{1}{c}\delta_h^{[2n]}v_h, \delta_h^{[2n]}v_h)_0 \geq \frac{1}{c_M} \parallel \delta_h^{[2n]}v_h \parallel_0^2$).

Remark 4.2 For the same value of n, the approximations $\delta_h^{[2n]}\chi_i$ and $\partial_h^{[2n]}\chi_i$ use the values of the function at the same points. Therefore the matrices $U_h^{[2n]}$ and $K_h^{[2n]}$ have the same bandwidth.

c - Order of the modified schemes in a homogeneous medium

Using $U_h^{[2n]}$ instead of M_h has no interest ! However, we can use both by introducing a convex combination

$$M_\alpha^{[2n]} = \alpha M_h + (1-\alpha)U_h^{[2n]}, \qquad 0 \leq \alpha \leq 1 \ , \tag{49}$$

and we can show that a proper choice of the parameter α allows us to gain two orders of accuracy when compared to the classical discretization, which corresponds to the particular choice $\alpha = 1$. Moreover, remark 4.2 shows that the matrix $\omega^2 M_\alpha^{[2n]} - aK_h^{[2n]}$ keeps the same bandwidth as in the classical scheme.

In a homogeneous medium the modified scheme can be rewritten as

$$\left\{ \begin{array}{ll} \dfrac{dw_h}{dz}(x_j,z) - \dfrac{i\omega}{c}\varphi_h(x_j,z) = 0 & (a) \\ \omega^2 \displaystyle\sum_i (M_\alpha^{[2n]})_{ji}\varphi_h(x_i,z) - \sum_i (K_h^{[2n]})_{ji}\,(a\varphi_h + bw_h)\,(x_i,z) = 0 & (b') \end{array} \right. , \tag{50}$$

and the truncation error still comes from the second equation and is defined as

$$E_h^{mod} = \omega^2 \sum_i (M_\alpha^{[2n]})_{ji}\varphi(x_i,z) - \sum_i (K_h^{[2n]})_{ji}\,(a\varphi + bw)\,(x_i,z) \ , \tag{51}$$

where (w,φ) is the exact solution of (20), assumed regular enough.

Proposition 4.1 *The modified scheme (43), with the matrix $M_\alpha^{[2n]}$ defined in (49) is of order $2n+2$ in a homogeneous medium with the choice*

$$\alpha = \alpha^{[2n]} = \frac{2n}{2n+1} \ . \tag{52}$$

Proof : we only mention here a lemma used in this proof, since it is useful, and refer to appendix C for the whole proof of this proposition. ■

Lemma 4.4 *The coefficients ν_p and μ_p satisfy the relation*

$$\mu_p = (2p-1)\nu_p, \quad 1 \leq p \leq n \ . \tag{53}$$

Proof : It is straightforward to check that the coefficients $(2p-1)\nu_p$ are solutions of system (45) and hence coincide with the μ_p ■.

Remark 4.3 From a practical point of view, relation (53) is very useful, since it is sufficient to compute a $2n$-order approximation of d/dx to get at the same time the $2n$-order approximation of the identity. Moreover, since coefficients μ_p satisfy a Vandermonde system, they are known explicitely, for all $1 \leq p \leq n$,

$$\mu_p = \frac{\displaystyle\prod_{m\neq p}(2m-1)^2}{\displaystyle\prod_{m\neq p}\left((2m-1)^2-(2p-1)^2\right)} = \frac{(2n!)^2(-1)^{p-1}}{2^{2(2n-1)}(2p-1)(n-p)!\,(n+p-1)!\,(n!)^2} \ , \tag{54}$$

and the expression for ν_p follows from (53).

Remark 4.4 - Interpretation of the modified schemes as an extension of the Claerbout's scheme. The classical Claerbout's scheme [11] is usually seen as a modification of the stiffness matrix $K_h^{[2]}$ in (33). The second-order approximation $K_h^{[2]}$ is replaced by $(I-\gamma h^2 K_h^{[2]})^{-1}K_h^{[2]}$ which is still a 2nd-order approximation and becomes of fourth-order with the value $\gamma = 1/12$. With $n=1$, the modified scheme corresponds to the classical Claerbout's scheme [11], with the relation $\gamma = (1-\alpha)/4$. The modification presented here plays on the mass matrix, but it can be interpreted as a modification on the stiffness matrix as in the Claerbout's scheme, by rewriting the modified mass matrix as

$$M_\alpha^{[2n]} = \left(I + (1-\alpha)(U_h^{[2n]} - M_h)M_h^{-1}\right) M_h \ ,$$

and multiplying the second equation in the approximate system by $\left(I + (1-\alpha)(U_h^{[2n]} - M_h)M_h^{-1}\right)^{-1}$ (assuming this is valid, which is true if $(1-\alpha) \parallel (U_h^{[2n]} - M_h)M_h^{-1} \parallel < 1$), in order to reobtain M_h for the mass term. The stiffness matrix is then modified and becomes

$$A_\alpha^{[2n]} = \left(I + (1-\alpha)(U_h^{[2n]} - M_h)M_h^{-1}\right)^{-1} K_h^{[2n]} \ . \tag{55}$$

The modified scheme can then be rewritten as follows

$$\left\{ \begin{array}{ll} \dfrac{dW_h}{dz} - i\omega M_h \Phi_h = 0 & (a) \\[2ex] \omega^2 M_h \Phi_h - A_\alpha^{[2n]}\,(a\Phi_h + bW_h) = 0 & (b') \end{array} \right. . \tag{56}$$

The corresponding value $\alpha^{[2]} = 2/3$ leads then to the well known $\gamma = 1/12$ choice and gives a fourth-order scheme with a matrix of bandwidth equal to 3 instead of 7 for the classical fourth-order scheme. The fourth-order modified scheme can thus be interpreted as the Claerbout's scheme and the higher order modified schemes as a generalization to higher orders of the Claerbout's scheme. For $n=2$ and $\alpha^{[4]} = 4/5$, we get a sixth-order scheme with a bandwidth equal to 7 instead of 11 for the classical sixth-order scheme.

The hypothesis to prove the equivalence between (56) and (43) that the matrix $\left(I + (1-\alpha)(U_h^{[2n]} - M_h)M_h^{-1}\right)^{-1}$ is invertible, is quite reasonable : it expresses that the mass matrix approximation that we use is not too far from the mass matrix!

4.2.5 Stability analysis

It is convenient to analyze the well-posedness of the schemes written in the form (56) which is very closed to the continuous equations so we have a guide for the proof of stability.

Proposition 4.2

- *If the matrix $A_\alpha^{[2n]}$ defined in (55) satisfies the following condition*

$$A_\alpha^{[2n]} V_h \bullet V_h \in \mathbb{R}, \qquad \forall V_h, \tag{57}$$

then the approximate problem (56) (or equivalently (43)) has a unique solution (W_h, Φ_h) that satisfies the energy conservation

$$\parallel W_h(z) \parallel = \parallel W_h(0) \parallel \qquad \forall z.$$

- *The classical schemes ($\alpha = 1$) satisfy condition (57).*
- *The modified schemes satisfy condition (57) in the case of a homogeneous medium.*

Proof : see appendix D. ■

In the proof of proposition 4.2, we end up with a quantity on the form $(c\partial_h c\delta_h \frac{1}{c}\delta_h t_h, \partial_h t_h)_0$ which has to be real. The key point in the homogeneous case is that c is constant and commutes with the other operators, so that this quantity is equal to $c(\partial_h \delta_h^2 t_h, \partial_h t_h)_0$ and since operators δ_h and ∂_h also commute, there is no difficulty to conclude. In the heterogeneous case, the velocity does not commute any more with the other operators. The main drawback of the modified scheme is therefore that the stability analysis through a priory energy estimates, that can be made for the classical schemes to show their well-posedness, fails in heterogeneous media, although it still applies in homogeneous media. However, from a numerical point of view, they seem to behave even better than the same order classical ones and this is also confirmed by the dispersion analysis.

In the following, for the dispersion analysis as well as for the numerical results, we only consider the 2n+2 order modified schemes obtained with the particular choice of $\alpha = \alpha^{[2n]}$.

4.3 Total discretization

Before closing this section, let us write the total discretization. System (43) can be rewritten after elimination of the auxiliary function as (when it is possible, i.e., if ω^2/a is such that the matrix $\frac{\omega^2}{a}M_\alpha^{[2n]} - K_h^{[2n]}$ can be inverted)

$$\frac{dW_h}{dz} = iC_hW_h, \qquad C_h = \omega M_h(\omega^2 M_\alpha^{[2n]} - aK_h^{[2n]})^{-1}bK_h^{[2n]}. \tag{58}$$

The discretization in depth is introduced as in section 4.1, thus if W_h^m denotes the approximation of $W_h(z^m)$, W_h^{m+1} is obtained from (30), substituting C with C_h. The algorithm for one iteration ($W_h^m \longrightarrow W_h^{m+1}$) is

• Compute the matrices at step z^m,

$$M^m = M_h(z^m) \quad , \quad M_\alpha^m = M_\alpha^{[2n]}(z^m) \quad , \quad K^m = K_h^{[2n]}(z^m).$$

• Introduce $K+1$ intermediary unknown vectors $(W_k^m)_{k=0,\dots,K}$

 $*$ $W_0^m = W_h^m$

 $*$ For $k = 1, \dots, K$, solve

$$S_k^m W_k^m = \bar{S}_k^m W_{k-1}^m \text{ , with } S_k^m = (\frac{\omega^2}{a}M_\alpha^m - K^m)(M^m)^{-1} + \bar{r}_k \frac{\omega \Delta z b}{a} K^m.$$

 $*$ $W_h^{m+1} = W_K^m$.

Notation - We will need a denomination to characterize the schemes we have presented. A scheme obtained using a $2n$-order discretization in x and a $2K$-order discretization in z will be called $(2nx - 2Kz)$-scheme. We also introduce the notation $(2nx_{class} - 2Kz)$-scheme and $(2nx_{mod} - 2Kz)$-scheme to distinguish between the classical and the modified $(2nx - 2Kz)$-order schemes.

5 Dispersion analysis in 2D

In this section we present the dispersion analysis of several numerical schemes presented in the previous section, i.e., the analysis of the quality of the schemes on the propagation of plane waves in homogeneous media. Since the application we have in mind for these equations concerns the migration in geophysics, we introduce here specific quantities commonly used in the geophysics community. So the dispersion analysis is done through the evaluation of what geophysicists call as "dip", a quantity defined below. The error on the dip (called the "dip error") is our criterion to quantify the accuracy of the schemes. This error depends on several parameters : two parameters related to the discretization (the number of discretization points per wavelength and $\Delta z/\Delta x$, the ratio between the step-sizes) ; one parameter related to the plane wave we modelize (the "apparent dip"). We will explain how to compute the numerical dip for the schemes presented in the previous section and give explicit expressions for several schemes that we study in more details. We then show the influence of these parameters on the dispersion error.

5.1 Some preliminaries

The dispersion analysis consists in analyzing the propagation of plane waves in a homogeneous medium, i.e., of waves on the form

$$v(x,z) = \exp -i(k_x x + k_z z), \tag{59}$$

where k_x and k_z are the components of the wave number. This solution corresponds in the time domain to an harmonic plane wave $V(x,z,t) = \exp -i(k_x x + k_z z - \frac{\omega}{c}t)$ where ω is the angular frequency. The slope of the equiphase lines in the (x,t) plane (time section) is called the **"apparent dip"**, $p_a = \dfrac{ck_x}{\omega} = \tan\theta_a$, and the slope of the equiphase lines in the (x,z) plane (depth section) is called the **"effective dip"**, $p_e = \dfrac{k_x}{k_z} = \tan\theta_e$. Functions (59) are solutions of the harmonic wave equation provided that k_x and k_z satisfy the classical dispersion relation

$$k_x^2 + k_z^2 = \frac{\omega^2}{c^2}. \tag{60}$$

From relation (60), one deduces a very simple relation between the apparent dip and the effective one for the wave equation, which is independent on the frequency

$$(p_e^2)^{wave} = \frac{p_a^2}{1-p_a^2}.$$

This analysis can also be applied to the paraxial equations. For the forty-five degree paraxial equation, v is governed by (20) and (21) which leads to a dispersion relation that can again be written as a relation between the apparent and the effective dips (still independent on the frequency)

$$\left(\frac{ck_z}{\omega}\right)^{cont} = 1 - \frac{bp_a^2}{1-ap_a^2} \iff p_e^{cont} = \frac{p_a}{1-\frac{bp_a^2}{1-ap_a^2}}. \tag{61}$$

We make a similar plane wave analysis for the schemes by looking for solutions on the form

$$v_j^m = \exp -i(k_x j\Delta x + k_z m\Delta z), \tag{62}$$

but this time the relation depends on the frequency, i.e., the numerical effective dip, p_e^{num}, given by this relation not only depends on p_a but also on ω. This is called the numerical dispersion. Of course, the weaker this dependence on the frequency is, the better the scheme is. We thus define a criterion that evaluates the quality of the scheme through its dispersion relation. The criterion we chose, called **the dip error** E, is the difference between the migrated plane wave given by the continuous paraxial equation and the migrated plane wave given by the scheme

$$\left\{\begin{array}{rcl} E &=& |\theta_e^{cont} - \theta_e^{num}| = \left|\arctan\left(\left(\frac{k_x}{k_z}\right)^{cont}\right) - \arctan\left(\left(\frac{k_x}{k_z}\right)^{num}\right)\right| \\ &=& \left|\arctan\left(\frac{p_a}{\left(\frac{ck_z}{\omega}\right)^{cont}}\right) - \arctan\left(\frac{p_a}{\left(\frac{ck_z}{\omega}\right)^{num}}\right)\right|, \end{array}\right. \tag{63}$$

The error clearly depends on the apparent dip and also on the discretization through the expression $\left(\frac{ck_z}{\omega}\right)^{num}$. Actually, it depends only on two parameters related to the discretization (see below) which are the dimensionless angular frequencies ζ and ζ_z relative to the x and z discretizations. It will also be useful to define the number of discretization points per wavelength in x, G, its inverse, $H = 1/G$, and the ratio between the step-sizes, r_{zx}. All these quantities are defined as follows

$$\zeta = \frac{\omega\Delta x}{c} \quad ; \quad r_{zx} = \frac{\Delta z}{\Delta x} \quad ; \quad G = \frac{2\pi c}{\omega \Delta x} \quad ; \quad \zeta_z = \frac{\omega\Delta z}{c} \; .$$

One has the following relations :

$$\zeta = 2\pi H \quad ; \quad \zeta_z = \zeta r_{zx} \quad ; \quad \Delta x \; k_x = \zeta \; p_a \; .$$

5.2 Dispersion relations of the numerical schemes

As explained above, the dispersion relation of the schemes is obtained by substituting solutions on the form (62) in the scheme. Practically, the dispersion analysis rests on the following remark. If Q is one of the matrices occurring in the scheme, Q acts as a diagonal matrix on every vector of plane wave type, i.e., if $\phi_j = e^{ijk_x\Delta x}$ then $Q\phi_j = \widehat{Q}(k_x\Delta x)\phi_j$. The scalar $\widehat{Q}(k_x\Delta x)$ is called the symbol of Q. For instance, if $Q\phi_j = \dfrac{\phi_{j+1} - 2\phi_j + \phi_{j-1}}{\Delta x^2}$, its symbol $\widehat{Q}(k_x\Delta x)$ is given by $-4\sin^2(k_x\Delta x/2)/\Delta x^2$. Now, from (30) and (58) and taking into account the Claerbout change of unknown ($W_h^m = e^{i\frac{\omega}{c}z^m}V_h^m$, V_h^m being the approximation of v), the scheme can be rewritten in the convenient form

$$V_h^{m+1} = e^{-i\frac{\omega\Delta z}{c}} \prod_{k=1}^{K} (I + \bar{r}_k \Delta z C_h)^{-1}(I + r_k\Delta z C_h)V_h^m. \tag{64}$$

From the previous remarks, it is straightforward to get the general dispersion relation for the scheme (64)

$$e^{i(\frac{\omega\Delta z}{c}-k_z\Delta z)} = \prod_{k=1}^{K} \frac{1+r_k\Delta z\widehat{C}_h}{1+\bar{r}_k\Delta z\widehat{C}_h} \equiv \frac{N_K(\Delta z\widehat{C}_h)}{\overline{N}_K(\Delta z\widehat{C}_h)}, \tag{65}$$

where $N_K(x)$ is the polynomial numerator, product of the numerators of each fraction (see (29)) and that yields the numerical dispersion relation

$$1-\left(\frac{ck_z}{\omega}\right)^{num} = \frac{2}{\zeta_z}\arctan\left(\frac{\Im m(N_K(\Delta z\widehat{C}_h))}{\Re e(N_K(\Delta z\widehat{C}_h))}\right). \tag{66}$$

$\Im m(N_K)$ (resp. $\Re e(N_K)$) being the imaginary part (resp. the real part) of N_K. This relation is valid for all the schemes introduced in the last section, N_K characterizes the discretization in z and the symbol $\widehat{C}_h$ characterizes the discretization in the lateral variable. This gives the expression of the numerical dip

$$\theta_e^{num} = \arctan\left(\frac{p_a}{1-\dfrac{2}{\zeta_z}\arctan\left(\dfrac{\Im m(N_K(\Delta z\widehat{C}_h))}{\Re e(N_K(\Delta z\widehat{C}_h))}\right)}\right). \tag{67}$$

The symbols $\widehat{C}_h$ for the $2n$-order classical ($\alpha = 1$) and the $(2n+2)$-order modified schemes ($\alpha = \alpha^{[2n]}$) can be expressed as

$$\Delta z\widehat{C}^{[2n]}_{h;\,class} = \zeta_z\frac{b\widetilde{K}^{[2n]}_h}{1-a\widetilde{K}^{[2n]}_h} \equiv \zeta_z\widetilde{C}_h \quad ; \quad \Delta z\widehat{C}^{[2n+2]}_{h;\,mod} = \zeta_z\frac{b\widetilde{K}^{[2n]}_h}{\widetilde{M}^{[2n]}_\alpha - a\widetilde{K}^{[2n]}_h} \equiv \zeta_z\widetilde{C}_h\ , \tag{68}$$

where the quantities $\widetilde{K}^{[2n]}_h$ and $\widetilde{M}^{[2n]}_\alpha$ depend only on ζ and p_a so that $\widetilde{C}_h = \widetilde{C}_h(\zeta, p_a\ ;\ a, b)$. Explicit expressions to compute θ_e^{num} from (67) and (68) for $K = 1, 2$ and $n = 1, 2, 3$ can be found in appendix E.

Remark 5.1 There is no difficulty to check, using Taylor expansions, that with a $(2nx - 2Kz)$-order scheme, one gets

$$\left(\frac{ck_z}{\omega}\right)^{num} = \left(\frac{ck_z}{\omega}\right)^{cont} + O(\Delta z^{2K}) + O(h^{2n})\ ,$$

i.e., as we expected, the dispersion error is of order $2K$ in z and $2n$ in x.

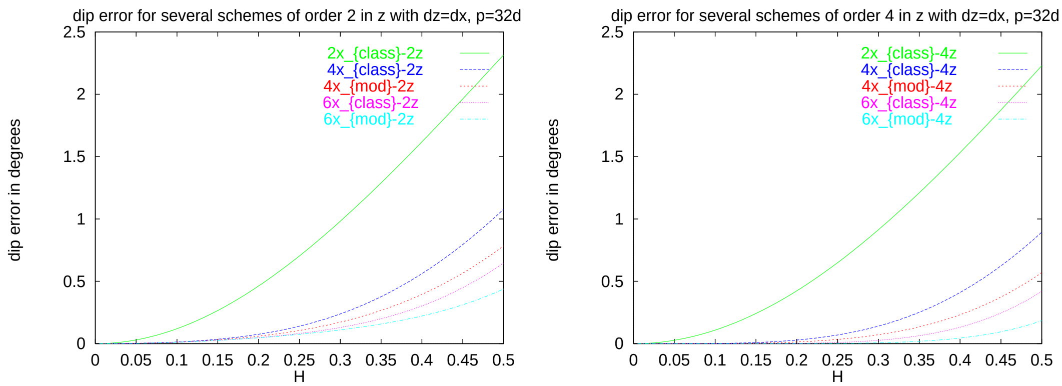
5.3 Comparison between several schemes

The dip error, computed from (61) and (66) (or equivalently (67)), depends on several parameters, $E = E(\zeta, r_{zx}, p_a; a, b)$. The two first parameters, ζ and r_{zx} are parameters of the scheme, they are directly related to the number of discretization points per wavelength for the discretization in x and for the discretization in z. The apparent dip p_a ($p_a < 1$) characterizes the propagating angle with the depth direction. One has to remember that the paraxial approximations are valid as long as the wave propagates in a direction close to the z direction, i.e., as long as p_a is small enough. For instance, for the forty-five degree approximation, the error between the paraxial approximation and the wave equation behaves as $O(p_a^6)$. The last parameters, a and b, are the coefficients of the rational fraction that defines the paraxial approximation. Here, we study only the approximations obtained with one fraction but the error for an approximation involving several fractions would be simply obtained as a sum of the errors for each fraction. This remark will be used to study the dispersion for the 3D paraxial equations with splitting. For the 2D case, taking one fraction means that we consider the forty-five degree approximation and thus make the analysis for $a = 1/4$ and $b = 1/2$.

For the sake of simplicity, we assume in the following that the step-sizes are the same in x and in z ($\Delta x = \Delta z$), i.e. $r_{zx} = 1$, so that the error becomes only a function of (ζ, p_a). We now compare the $(2nx - 2Kz)$ classical and modified schemes for $K = 1, 2$ and $n = 1, 2, 3$. One could think that with a sixth-order discretization in x it would be better to use also a sixth-order discretization in z. However the results obtained with the sixth-order discretization in z did not show any improvement compared with the fourth-order, that is why it is not presented.

5.3.1 Comparison between classical and modified schemes

Fig. 9 (a) and (b) represent the dip error $E(\zeta = 2\pi H, p_a)$ for several discretizations in x and a fixed discretization in z (2nd-order in (a), 4th-order in (b)). The value of the apparent dip is fixed to $p_a = \tan(32°) \approx 0.6$ (but the conclusions remain valid for p_a describing the interval $[0, 1]$). For the second-order as well as for the



(a) **Second-order schemes in** z (b) **Fourth-order schemes in** z

Figure 9: **Comparison between classical and modified schemes**

fourth-order z-discretization, Figures 9(a) and (b) show the same classification for the schemes :

$$\left\{\begin{array}{ccccccccc} \boldsymbol{2x_{class}} & < & \boldsymbol{4x_{class}} & < & \boldsymbol{4x_{mod}} & < & \boldsymbol{6x_{class}} & < & \boldsymbol{6x_{mod}} \\ \boldsymbol{bw} = 3 & & \boldsymbol{bw} = 7 & & \boldsymbol{bw} = 3 & & \boldsymbol{bw} = 11 & & \boldsymbol{bw} = 7\ , \end{array}\right.$$

where $S_1 < S_2$ means that the scheme S_2 is better than the scheme S_1 from the dispersion point of view. Under each scheme, we have indicated the bandwidth $\boldsymbol{bw}$ of the linear system to be solved. Consequently, for the same order of accuracy the modified schemes give better results than the classical ones. This was not predictable and is all the more interesting because the modified schemes have been designed in such a way that they also consume less computational time (see the bandwidths).

5.3.2 Comparison between second and fourth-order discretizations in z

When comparing Figures 9 (a) and (b), it can be seen that there is almost no improvement by increasing the order of discretization in z with the classical second-order scheme in x. On the other hand, for the other discretizations in x (i.e., of order greater than 4) the gain is important especially for small values of H, i.e. for a large number of points per wavelength (see Fig 10). In any case, one should remember that the cost of the 4th-order scheme is around twice the cost of the 2nd-order one. Concerning the price we should pay, we address the following question: is it worth using a fourth-order depth discretization rather than using the second-order one with a step divided by two and the double of iterations ? (since the price should be the same). To answer this question, we have compared the dip errors obtained with the 2nd-order z-discretization for the ratio $r_{zx} = 1/2$ and with the 4th-order z-discretization for the ratio $r_{zx} = 1$, and this for three different discretizations in the lateral variable ($2x_{class}$, $4x_{mod}$ and $6x_{mod}$). The conclusion is always the same, for the same computational cost, the 4th-order z-discretization is always more accurate than the 2nd-order one. Actually, to get a better accuracy with the 2nd-order z-discretization compared to the 4th-order one, one should divide the step-size in z by 12! it would thus be 6 times more expensive. In conclusion, for a better accuracy, we should better use the fourth-order z-discretization than decrease the step-size.

5.3.3 Comparison between the modified schemes $4x_{mod} - 4z$ and $6x_{mod} - 4z$

In the last two sections, the comparison was done for a fixed discretization either in z or in x. We now wonder if to get a better accuracy it is worth using a higher-order discretization in x, say the 6th-order modified scheme, with the 2nd-order z-discretization or rather the 4th-order discretization in both variables ? The dip error for these two schemes is plotted in Fig 10, for the value $p_a = \tan(32°)$. The two curves cross each other at

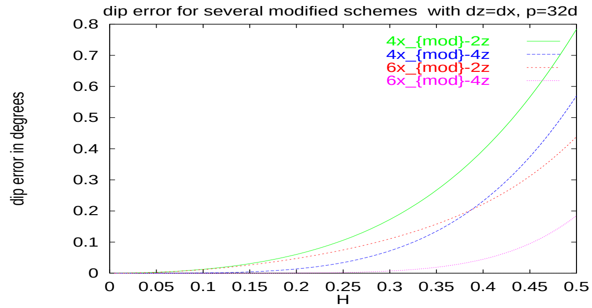


Figure 10: Comparison between several modified schemes, $p_a = \tan(32°)$

$H \approx 0.38$ which corresponds to about 2.6 points per wavelength. This conclusion is still valid for other values of the dip, except that the value of the cross point is not the same anymore, and varies from $H \approx 0.38$ to $H \approx 0.42$, see Figure 11. This corresponds to a value of G between 2.3 and 2.6 points per wavelength. Therefore it appears that when we have a large enough number of points per wavelength (larger than 2.5, which is not so large...), it is better to use the 4th-order scheme $4x_{mod} - 4z$, whereas for a poor mesh (less than 2.5 points per wavelength) the scheme $\boldsymbol{6x_{mod} - 2z}$ is better. However, even if one of the main advantage of paraxial equations, compared to the wave equation, is to need a lower number of points per wavelength, it is certainly not realistic to use less than 2.5 points per wavelength with the hope of getting an accurate solution!

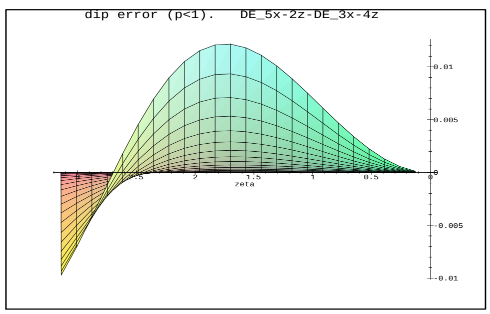


Figure 11: Comparison between schemes $\boldsymbol{6x_{mod} - 2z}$ and $\boldsymbol{4x_{mod} - 4z}$ for several values of the dip

6 Application to the 3D solution

6.1 Algorithm

We come back to the approximation of the 3D field u solution of (15) (or equivalently to the related seismic field v). For the sake of simplicity, we assume here that the paraxial approximation uses only one fraction per direction of splitting. Assume the approximate solution V_h^m (here the subscript h only means that it is the approximate solution), at step z_m, is known, the algorithm to compute the solution V_h^{m+1} is summarized as follows: introduce $N_D + 1$ intermediate unknowns, $(W^{m,j})_{j=0,\ldots,N_D}$,

1. **transport term**
$$W^{m,0} = e^{\frac{i\omega \Delta z}{c^m}} V_h^m \ .$$

2. **Splitting in the N_D directions**

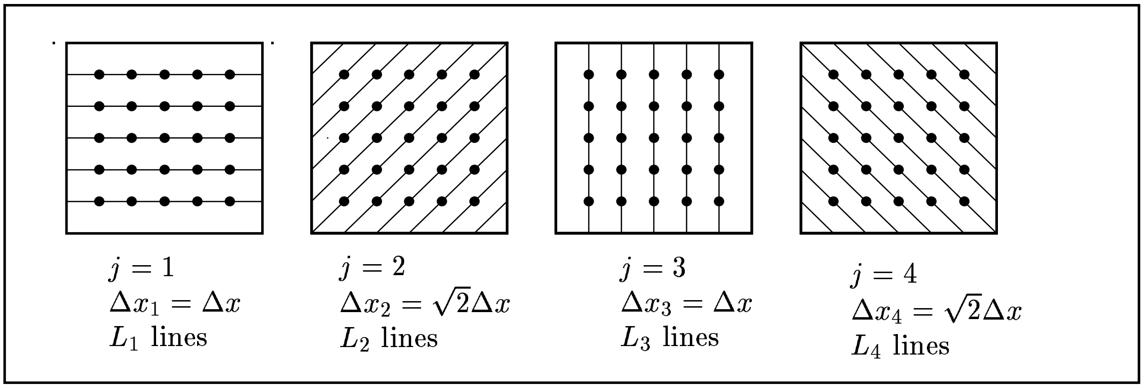


Figure 12: Computational grid for the 45 degree approximation using 4 directions of splitting

For each direction $1 \leq j \leq N_D$, on each mesh-line in this direction, $1 \leq l_j \leq L_j$, $W^{m,j}$ is determined through the solution of the 2D paraxial equation (19) on this line, whose semi-discretization (using the modified schemes) is

$$(69) \qquad \left\{ \begin{array}{l} \dfrac{dW_j^m}{ds} = iC_h^{j,m} W_j^m, \quad 0 \leq s \leq \Delta z \\ W_j^m(0) = W^{m,j-1} \\ W^{m,j} \equiv W_j^m(\Delta z) \end{array} \right. ,$$

with $C_h^{j,m} = \omega M_h(\omega^2 M_\alpha^{[2n]} - a_j K_h^{[2n]})^{-1} b_j K_h^{[2n]}$. The matrices depend on the depth index m, on the direction j and on the considered line l_j. When using 4 directions, as illustrated in figure 2, all the systems to solve do not have the same size. The step-size also depends on the direction: it is larger in the diagonal directions. To avoid too much dispersion coming from these directions, it is thus important to choose the number of discretization points with respect to the largest step-size. The total discretization of (69) follows the algorithm described in section 4.3: introduce $K+1$ intermediary unknown vectors $(\widetilde{W}_j^m)_{k,\ k=0,\dots,K}$

* $(\widetilde{W}_j^m)_0 = W^{m,j-1}$
* For $k = 1, \dots, K$, $(\widetilde{W}_j^m)_k$ is solution of the linear system

$$S_k^{m,j}(\widetilde{W}_j^m)_k = \bar{S}_k^{m,j}(\widetilde{W}_j^m)_{k-1} \text{ , with } S_k^{m,j} = (\frac{\omega^2}{a_j} M_\alpha^{[2n]} - K_h^{[2n]})(M_h)^{-1} + \bar{r}_k \frac{\omega \Delta z b_j}{a_j} K_h^{[2n]}.$$

* $W^{m,j} \equiv (\widetilde{W}_j^m)_K$.

3. **Update of the solution**

$$V_h^{m+1} = W^{m,N_D} \ .$$

For the practical implementation, we introduced a local numerotation for the nodes on each line in each direction, and a correspondance between the local numerotation and the global numerotation in the mesh.

6.2 Optimal choice for the coefficients of the 3D splitted forty-five degree paraxial equation from the dispersion point of view

In this section, we extend the dispersion analysis done in 2D to 3D paraxial equations. We introduce the quantities

$$\left\{ \begin{array}{l} \vec{p}_a = (\dfrac{ck_x}{\omega}, \dfrac{ck_y}{\omega}) = p_a(\cos\beta, \sin\beta) \quad \text{ with } \quad p_a = \| \vec{p}_a \| \\ \vec{p}_e = (\dfrac{k_x}{k_z}, \dfrac{k_y}{k_z}) \quad ; \quad p_e = \| \vec{p}_e \| = \tan\theta_e = \dfrac{p_a}{\frac{ck_z}{\omega}} \ . \end{array} \right.$$

The dispersion relation of the 3D paraxial equations takes into account each fraction

$$(70) \qquad \left(\frac{ck_z}{\omega} \right)^{cont} = 1 - \sum_{j=1}^{N_D} \sum_{\ell=1}^{L} \frac{b_j^\ell (\vec{p}_a . n_j)^2}{1 - a_j^\ell (\vec{p}_a . n_j)^2}.$$

In the following, we restrict ourselves to the family of forty-five degree approximations using 4 directions for the splitting and 1 fraction per direction, that has been described in section 3 (see relations (12)). Similar analysis could be applied for other approximations, involving either other directions for the splitting or more fractions.

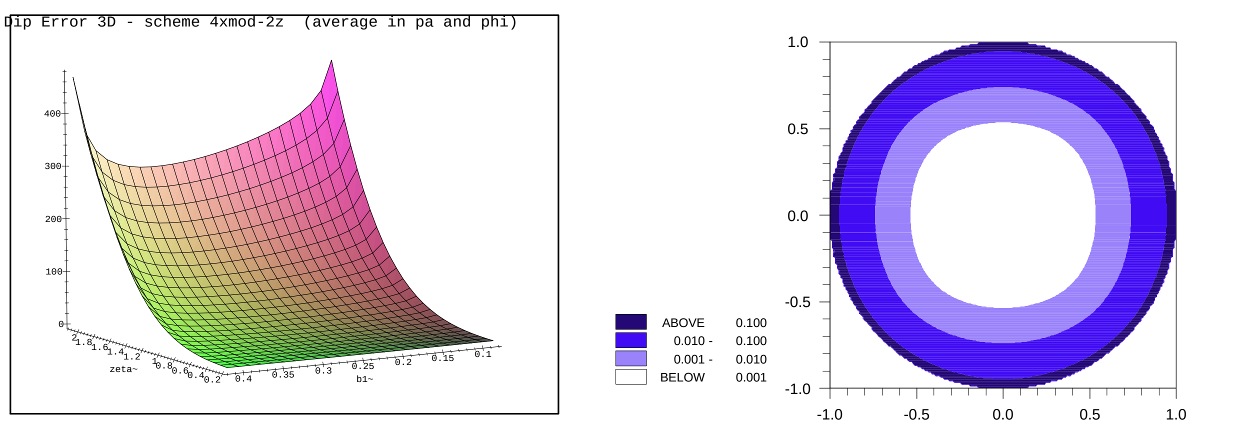
Concerning the scheme, the difference with the 2D case results of course from this sum. At each step, we thus have to solve in each direction a 2D problem which corresponds to the following scheme :

$$V_h^{m+1} = e^{-i\frac{\omega \Delta z}{c}} \prod_{j=1}^{N_D} \prod_{k=1}^{K} (I + \bar{r}_k \Delta z C_h^j)^{-1} (I + r_k \Delta z C_h^j) V_h^m \ , \tag{71}$$

where C_h^j represents the operator in the j-th direction (since we consider here a homogeneous medium, the operators do not depend on m). The corresponding dispersion relation can be written as

$$1 - \left(\frac{ck_z}{\omega}\right)^{num} = \frac{2}{\zeta_z} \sum_{j=1}^{N_D} \arctan\left(\frac{\Im(N(\zeta_z \widetilde{C}_h^j))}{\Re(N(\zeta_z \widetilde{C}_h^j))}\right) . \tag{72}$$

In (72), the symbol $\widetilde{C}_h^j$ is related to function $\widetilde{C}_h(\zeta, p_a \ ; \ a, b)$ defined in section 5.2, by $\widetilde{C}_h^j = \widetilde{C}_h(\zeta_j, \vec{p}_a.n_j \ ; \ a_j, b_j)$ where ζ_j denotes the dimensionless angular frequency in the j-th direction. From now on we assume that the spatial mesh is composed with squared elements ($\Delta x_1 = \Delta x_2$) thus we have $\zeta_1 = \zeta_4 = \dfrac{\omega \Delta x_1}{c} \equiv \zeta$ in the axis directions and $\zeta_2 = \zeta_3 = \sqrt{2}\zeta$ in the diagonal directions. The dip error is still defined by (63) and depends here on $E = E(\zeta, r_{zx} = 1, p_a, \beta \ ; \ b_1)$ since the other coefficients a_j and b_j are related to b_1 through relations (12). As recalled in section 3, the values of parameter b_1 proposed in [14] were chosen according to criteria based on the error $e(\kappa_1, \kappa_2)$, i.e., the error between the wave equation symbol and the paraxial equation one. Since the dip error also depends on this parameter b_1, we determine here the value that minimizes this error.



(a) **Optimal choice for the parameter b_1 and for the scheme $4x_{mod} - 2z$ (average with respect to p_a and ϕ, $r_{zx} = 1$)**

(b) **Error for the optimal value** $b_1 = 0.3$ ($\Delta x_2 = \Delta x_1$)

Figure 13: Optimal value for b_1 and corresponding error

In figure 13-(a), we represent the L^1 norm of the error, with respect to $\beta \in [0, 2\pi]$ and $p_a \in [0, 0.9]$ (i.e., $p_a \leq \tan 42°$),

$$\int_0^{2\pi} \int_0^{0.9} E(\zeta, 1, p_a, \beta; \ b_1) dp_a d\beta \ ,$$

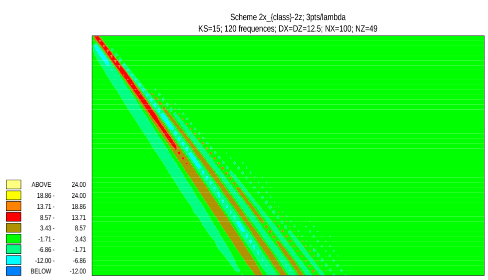
for ζ varying in $[2\pi/30, 2\pi/3]$ (i.e., from 30 to 3 points per wavelength in the axis directions) and b_1 varying in $[1/12, 5/12]$ and for the scheme $\boldsymbol{4x_{mod} - 2z}$. The optimal choice for b_1 is $b_1 \approx 0.3$, which is not far from the maxi isotropic corresponding to $b_1 = 0.25$). For the other schemes, and for a ratio r_{zx} between 0.5 and 2, the optimal value remains around 0.3. The criterium used to determine this value is based on the difference between the scheme and the continuous paraxial equation but it does not guaranty any more that the paraxial approximation approximates well the wave equation. In figure 13-(b), we represent the error $e(\kappa_1, \kappa_2)$ for this optimal value and we check that the approximation is quite good and seems even better than the maxi isotropic one (see figure 3-(a)), especially in the diagonal directions $\kappa_1 = \kappa_2$.

7 Numerical experiments

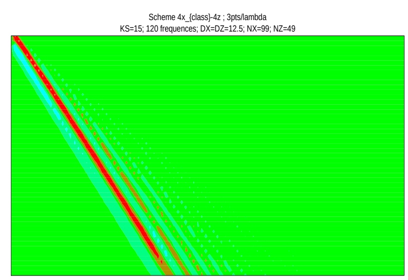
We now illustrate the method with several numerical experiments done in the time domain. In each case, we specify the surface data $R(x,t)$ in the time domain. Since in practice we solve the problem for each frequency and recover the transient result through a Fourier transform, we also indicate the value of the cutoff frequency, F_c. The computational domain is 1250 m in each of the horizontal directions, and 625 m in the vertical direction. The grid sizes are $h = \Delta z = 12.5$ m. We handle 120 equidistributed frequencies.

7.1 Migration of a straight-line reflector in a 2D homogeneous medium

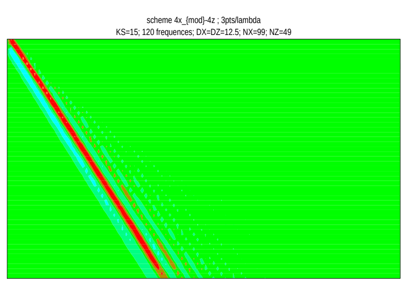
The surface data is given by $R(x,t) = f(t - \tau(x))$ where $\tau(x) = \dfrac{p_a}{c} x$ and f is the derivative of a Gaussian function $f(t) = \dfrac{d}{dt}\left(g_{\omega_S}(t)\right) = \dfrac{d}{dt}\left(\exp(-\omega_S^2 t^2/4)\right)$. The source frequency is taken to be $F_S = 30$ Hz and the cutoff frequency to $F_C = 78$ Hz. Theoretically, i.e. with the exact wave equation, we should observe a migrated image located around the straight line of equation $z = p_e x$ with $p_e = p_a/\sqrt{1 - p_a^2}$. The dispersion analysis showed that the error increases with the apparent dip and also with the inverse of the number of points per wavelength. We consider here a quite severe test with a high value of the dip: $p_a = 0.8$ (which corresponds to a reflector located on a straight line of slope $\approx 53°$) and a very small number of points (less than 3 for the source frequency and less than 2 for the cutoff frequency). Figure 14(a) represents the solution obtained with the classical



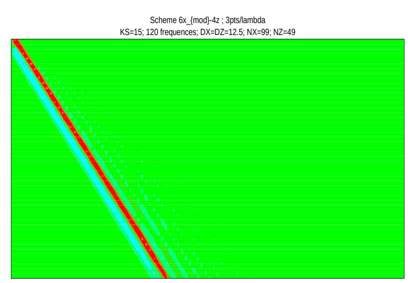
(a) scheme $\boldsymbol{2x_{class} - 2z}$



(b) scheme $\boldsymbol{4x_{class} - 4z}$



(c) scheme $\boldsymbol{4x_{mod} - 4z}$



(d) scheme $\boldsymbol{6x_{mod} - 4z}$

Figure 14: **Migration of a straight line reflector**

2nd-order scheme, $\boldsymbol{2x_{class} - 2z}$ and we observe an important dispersion. We then represent in Figures 14 (b) and (c) the classical and modified fourth-order schemes ($\boldsymbol{4x_{class} - 4z}$ and $\boldsymbol{4x_{mod} - 4z}$). The dispersion is already attenuated specially with the modified scheme. This confirms the conclusions of the dispersion analysis. Finally, with the use of the modified sixth-order scheme in x and 4th-order in z, $\boldsymbol{6x_{mod} - 4z}$, the numerical dispersion has almost disappeared (see Figure 14(d)). We do not show the results with 6th-order schemes in z as there is no significant further improvement.

7.2 Migration of a filtered point source in a 2D heterogeneous medium

The initial condition at $\boldsymbol{z = 0}$ is a filtered point source given as

$$R(x,t) = \frac{d^2}{dt^2}\left(g_{\omega_S}(t - T_S)\right)\delta(x - x_S) \star \mathcal{F}^{-1}_{x,t}\left(1_{|ck_x|<|\omega|}\right), \tag{73}$$

where g_{ω_S} is the Gaussian function defined as in section 7.1. This corresponds to the second time derivative of a Gaussian function in time multiplied by a filtered point source function, which means that the waves corresponding to evanescent modes for the wave equation have been eliminated. The inverse Fourier transform with respect to x can be explicited as

$$\mathcal{F}^{-1}_{x}(1_{|ck_x|<|\omega|})(x) = \frac{2}{\mid x \mid}\sin(\frac{\omega \mid x \mid}{c}).$$

The point source is located at the surface, in the center of the computational domain. This source thus contains all the values of the dip and therefore is more delicate from the dispersion point of view. The simulation is done in a smoothly varying velocity medium (see Figure 15). The central frequency of the source is $F_S = 28$ Hz and the cutoff frequency $F_C = 76$ Hz. For heterogeneous media the stability of the modified schemes has not been

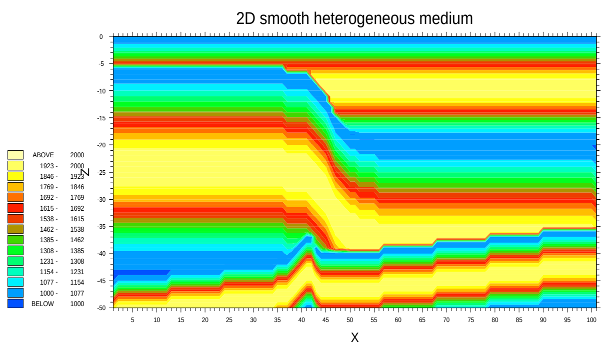


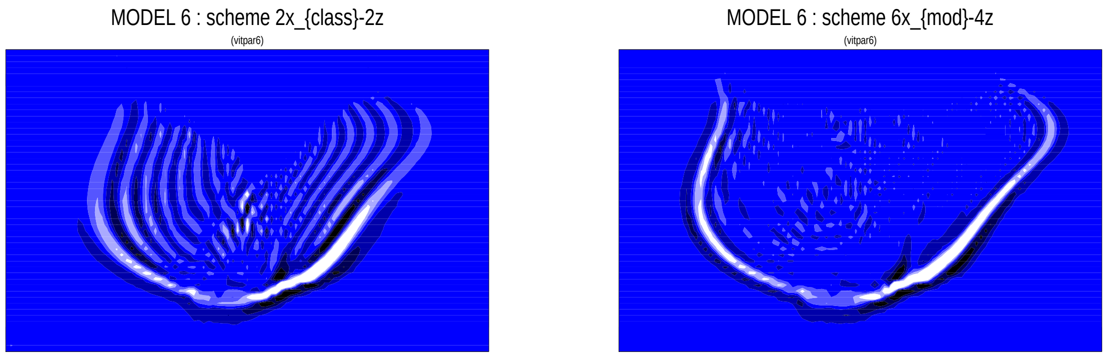
Figure 15: **2D smooth heterogeneous medium**

proved yet. However, the numerical experiments show that they still give good results and we observe again a very good improvement from the dispersion point of view compared to the classical 2nd-order scheme (see Figure 16).

In a second experiment, we shift the location of the source close to the left boundary of the domain. If we use Dirichlet boundary conditions at the lateral boundary, the migrated image shows a strong reflexion (see Figure 17(a)). In order to remove this spurious reflexion, we have adapted the PML technique proposed by Collino (1995) to our higher order schemes. The method has been described in section 3.2. In this example σ is zero in the domain of interest and piecewise constant in a vertical layer of five step size. The result obtained with the scheme $\boldsymbol{4x_{mod} - 4z}$ is depicted in Figure 17. It can be seen that the spurious reflection has disappeared. The extra cost due to the PML is negligible as only five extra nodes have been added to the computational domain. As a result, the PML technique can be extended to heterogeneous media and higher order schemes.

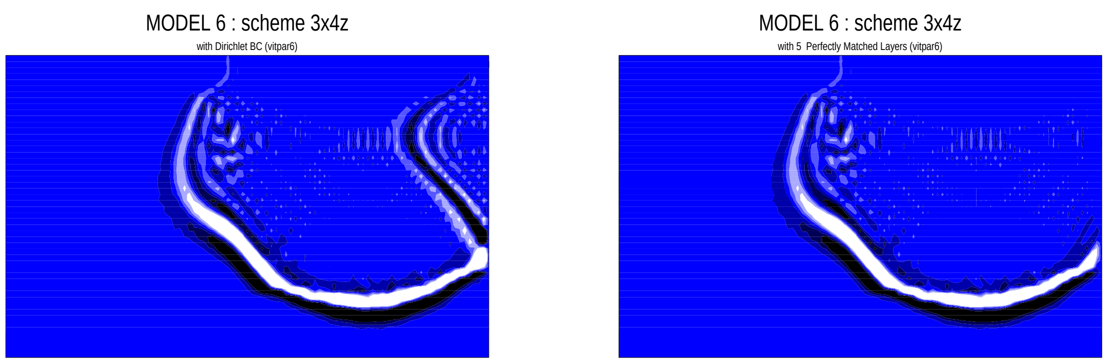
7.3 Migration of a filtered point source in a 3D homogeneous medium

This simulation is done in a 3D homogeneous medium with velocity equal to 1000m/s. We use a forty-five degree paraxial equation with 4 directions and 1 fraction per direction. The degree of freedom b_1 has been chosen



(a) scheme $\boldsymbol{2x_{class} - 2z}$ (b) scheme $\boldsymbol{6x_{mod} - 4z}$

Figure 16: **Migration of a filtered point source in a 2D smooth heterogeneous medium**



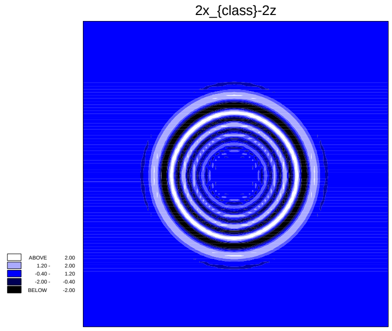
(a) Dirichlet BC (b) 5 PMLs

Figure 17: **Migration of a filtered point source in a 2D smooth heterogeneous medium. Scheme $\boldsymbol{4x_{mod} - 4z}$**

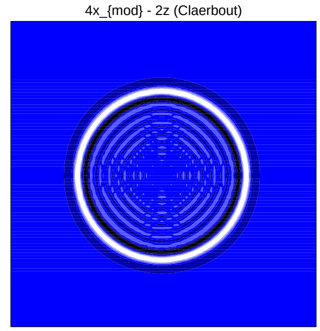
following the dispersion analysis of section 6.2 and taken equal to $b_1 = 0.3$. The equation is finally characterized by the coefficients $a_1 = a_4 = 0.27$, $b_1 = b_4 = 0.3$ in the directions x_1 and x_2 and $a_2 = a_3 = 0.41$, $b_2 = b_3 = 0.2$ in the diagonal directions. Again we consider a filtered point source, the only difference with the 2D case comes from the inverse Fourier transform which becomes in 3D

$$\mathcal{F}^{-1}_{x_1,x_2}(1_{|ck|<|\omega|})(x_1,x_2) = \frac{1}{2\pi}\left|\frac{\omega}{cx}\right| J_1\left(\frac{|\omega x|}{c}\right),$$

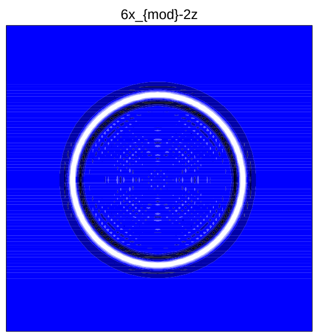
where J_1 denotes the Bessel function. The central frequency of the source is $F_S = 20$ Hz and the cutoff frequency $F_C = 50$ Hz. The number of points per wavelength is around 4 for the central frequency along the x_1 direction, but only ≈ 3 along the diagonal, and less than 2 for the cutoff frequency. Except for the 2nd-order, we only present here the results obtained with modified schemes since it is now clear from the dispersion analysis as well as from the previous results, that for the same accuracy, they are better and less expensive than the classical ones. We represent in Figures 18 the sections at a fixed depth. Also one should notice in the homogeneous case the quite good isotropy obtained with these new paraxial equations despite of the introduction of particular directions used for the splitting.



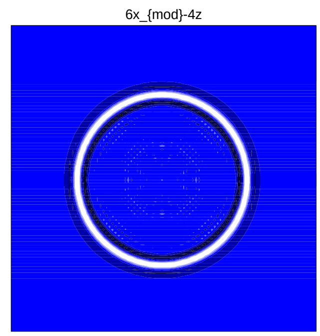
(a) **Classical** 2nd-order in x and 2nd-order in z



(b) **Modified** 4th-order in x and 2nd-order in z (Claerbout)



(c) **Modified** 6th-order in x and 2nd-order in z



(d) **Modified 6th-order** scheme (4th-order in z)

Figure 18: **Filtered point source in a 3D homogeneous medium**

7.4 Migration of a filtered point source in a 3D heterogeneous medium

The last simulation is done in a smooth varying velocity medium. This medium has been obtained by applying an operation (rotation with respect to the origin + homothetie) to the 2D model represented in figure 19. In figure 20 we represent again the sections for two fixed depth ($z = 20$ in (a), (b) and (c) and $z = 30$ in (d) and (e)). The improvement on the dispersion using a higher order scheme is still quite good in this heterogeneous medium (compare figures (a) and (c)). The PML technique is also used here. Although the extra cost is a bit higher and the absorption with only 6 layers is not so perfect than in the 2D case, the results compared to the Dirichlet Boundary Conditions are still quite good (compare figures (a) and (b) at depth $z = 20$ and figures (d) and (e) at depth $z = 30$).

8 Comparisons of the costs

We now compare the schemes from their computational cost. The systems have been with a LU factorization. The computations are done on a DEC Alpha work station with a 275 MHz cpu.

8.1 *Cpu* times of the experiments done in 2D

In order to compare several schemes, we take as a reference *cpu* time the one obtained for the classical second order scheme $\boldsymbol{2x_{class} - 2z}$, and we call it cpu_0. The indicated values in table 1 are averages obtained from

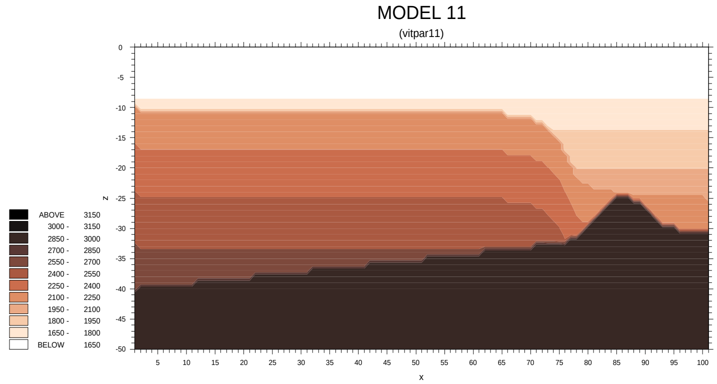
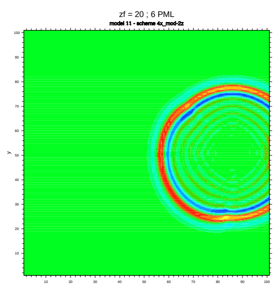
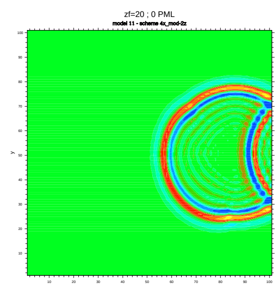


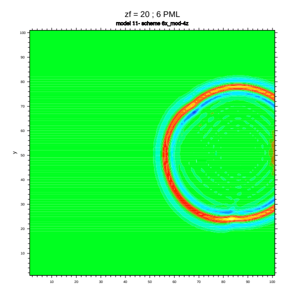
Figure 19: (x, z) **slice of a 3D smooth heterogeneous medium**



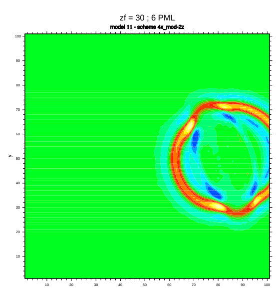
(a) $z = 20$, Scheme $\boldsymbol{4x_{mod} - 2z}$ (Claerbout) with 6 PML



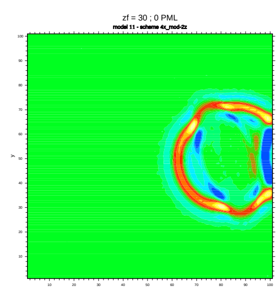
(b) $z = 20$, Scheme $\boldsymbol{4x_{mod} - 2z}$ with Dirichlet BC



(c) $z = 20$, Scheme $\boldsymbol{6x_{mod} - 4z}$ with 6 PML



(d) $z = 30$, Scheme $\boldsymbol{4x_{mod} - 2z}$ with 6 PML



(e) $z = 30$, Scheme $\boldsymbol{4x_{mod} - 2z}$ with Dirichlet BC

Figure 20: **Filtered point source in a 3D heterogeneous medium**

a significant number of experiments. For each scheme, we indicate two values : the first one is the average of the *cpu* time "per node" , i.e., the *cpu* time divided by the number of points ($N_x N_z$) and by the number of frequencies (N_f) and the second one represents the ratio between the *cpu* time and the reference cpu_0.

As announced, the modified schemes are much less expensive than the classical ones. It can be noticed that the cost of the fourth-order schemes in z is less than twice the cost of the second-order ones, as we expected ($cpu(\boldsymbol{4x_{mod} - 4z})/cpu(\boldsymbol{4x_{mod} - 2z}) \simeq 1.3$ and $cpu(\boldsymbol{6x_{mod} - 4z})/cpu(\boldsymbol{6x_{mod} - 2z}) \simeq 1.6$).

2D

	cpu	*cpu*/*cpu*$_0$
$2x_{class} - 2z$	5.52e-06	1.
$4x_{class} - 2z$	9.89e-06	1.79
$4x_{mod} - 2z$	5.7e-06	1.03
$4x_{mod} - 4z$	8.8e-06	1.6
$6x_{class} - 2z$	1.48e-05	2.68
$6x_{mod} - 2z$	1.05e-05	1.9
$6x_{mod} - 4z$	1.63e-5	2.95

3D

	cpu	*cpu*/*cpu*$_0$	3D/2D
$2x_{class} - 2z$	1.5e-05	1.	2.71
$4x_{class} - 2z$	2.8e-5	1.86	2.83
$4x_{mod} - 2z$	1.7e-5	1.1	2.98
$4x_{mod} - 4z$	2.5e-5	1.66	2.84
$6x_{class} - 2z$	4.3e-5	2.86	2.9
$6x_{mod} - 2z$	3.e-5	2.	2.85
$6x_{mod} - 4z$	5.3e-5	3.53	3.25

Table 1: **Comparison of *cpu* times. Left: 2D - Right: 3D**

8.2 *Cpu* times of the experiments done in 3D

The comparison for the 3D case is presented in 1-right. This time, the cost per node is around 2.9 times the one in the 2D case. The other conclusions concerning the comparisons between the schemes are about the same than in 2D, and we summarize it in the next section.

8.3 Conclusions on the *cpu* time

In the dispersion analysis, we have seen that the modified schemes have a better behavior than the classical ones. Their price, for a same accuracy, is also much less expensive. In the following, we thus focus on these modified schemes. These four schemes ($4x_{mod} - 2z$, $4x_{mod} - 4z$, $6x_{mod} - 2z$ and $6x_{mod} - 4z$) have been compared in section 5.3.3 from the dispersion point of view and we now indicate the ratios of their respective costs in order to precise what is "more expensive"...

We summarize the results of the 2D and the 3D cases. In table 2, we represent the cost ratios of these schemes, more precisely, the value indicated in the table represents the cost of the scheme of the corresponding column divided by the cost of the scheme of the corresponding line (the upper value is for the 2D case and the lower for the 3D case). If we want a better accuracy than the $4x_{mod} - 2z$ scheme, we recall from the dispersion analysis that for a large enough number of discretization points per wavelength, the $4x_{mod} - 4z$ scheme behaves better than the $6x_{mod} - 2z$ one, and we now see that it is also cheaper. On the contrary, for a small number of discretization points per wavelength, the $6x_{mod} - 2z$ scheme gives better results and is only around 1.2 times more expensive than the $4x_{mod} - 4z$ one. Finally, the most accurate scheme, $6x_{mod} - 4z$, which is also the most expensive one, is around 1.5 times more expensive than the previous, $6x_{mod} - 2z$.

		$4x_{mod} - 2z$	$4x_{mod} - 4z$	$6x_{mod} - 2z$	$6x_{mod} - 4z$
$4x_{mod} - 2z$	3D	1.	1.5	1.8	2.8
	2D	1.	1.5	1.8	3.2
$4x_{mod} - 4z$	3D	0.64	1.	1.1	1.8
	2D	0.66	1.	1.2	2.1
$6x_{mod} - 2z$	3D	0.54	0.84	1.	1.5
	2D	0.55	0.83	1.	1.7
$6x_{mod} - 4z$	3D	0.34	0.54	0.64	1.
	2D	0.31	0.47	0.56	1.

Table 2: **Ratio of the costs**

9 Conclusion

We have presented new higher-order numerical schemes to approximate either the 2D paraxial equations or the splitted 3D paraxial equations. These schemes have been designed to fulfill two main requirements : the extension to heterogeneous media and the treatment of the lateral boundaries. The dispersion analysis as well as the numerical results show that the numerical dispersion occurring with the classical second-order schemes can be considerably attenuated with these schemes, particularly with the modified higher-order schemes, even with coarse discretization grids. Thanks to the dispersion analysis for the 3D forty-five degree approximations, we have been able to determine the optimal equation from the dispersion point of view.

Furthermore the dispersion analysis gives a better idea of the behavior of each scheme from the dispersion point of view. In particular, for given parameters (apparent dip and number of points per wavelength) it gives a classification between the schemes and thus helps for choosing the "best" one. In practice, the time domain solution comes from a numerical Fourier transform of the frequency domain solution, which means that we deal with a more or less large number of frequencies. A possible strategy, that takes into account the informations given by the dispersion analysis, could be to adapt the choice of the scheme to the handled frequency. In order to really compare the different schemes, it remains to make the analysis of their cost, including the cost of construction of the matrices in heterogeneous media, since this cost seems to be not negligible at all.

One should also compare the efficiency of these equations with the full paraxial equation. This is work in progress and will be the subject of a forthcoming paper.

APPENDIX

A Proof of Lemma 4.2

We assume the exact solution (w,φ) to be very regular, and we use again the auxiliary function $\psi = a\varphi + bw$. The truncation error (41) can be rewritten, using equation (20)-(b) as

$$E_h^{class} = -c\frac{\partial^2\psi}{\partial x^2}(x_j,z) - \sum_i (K_h^{[2n]})_{ji}\psi(x_i,z)\ .$$

The matrix $K_h^{[2n]}$ in the homogeneous case is (we omit the index $[2n]$ in the following)

$$(K_h)_{ij} = c(\partial_h\chi_i, \partial_h\chi_j) = c(\partial_h^\star\partial_h\chi_i, \chi_j) = -c(\partial_h^2\chi_i,\chi_j)$$
$$\Longrightarrow \sum_i (K_h)_{ji}\psi(x_i,z) = -c\sum_i(\partial_h^2\chi_i,\chi_j)\psi(x_i,z)\ .$$

Since χ_i is in H_h, $\partial_h\chi_i$ is "shifted" in $H_h^{1/2}$ and $\partial_h^2\chi_i$ is back in H_h and can then be decomposed on the basis

$$\partial_h^2\chi_i(x) = \sqrt{h}\sum_m \partial_h^2\chi_i(x_m)\chi_m(x) \quad\Longrightarrow\quad (\partial_h^2\chi_i,\chi_j) = \sqrt{h}\partial_h^2\chi_i(x_j)\ .$$

Finally, we get

$$\sum_i (K_h)_{ji}\psi(x_i,z) = -c\sqrt{h}\sum_i \partial_h^2\chi_i(x_j)\psi(x_i,z) \equiv -c\partial_h^2(\pi_h\psi)(x_j,z)\ ,$$

the second equality coming from the definition of π_h, (32). The error becomes therefore

$$\begin{aligned} E_h^{class} &= -c\frac{\partial^2\psi}{\partial x^2}(x_j,z) + c\partial_h^2(\pi_h\psi)(x_j,z) \\ &= c\left[-\frac{\partial^2\psi}{\partial x^2}(x_j,z) + \partial_h^2\psi(x_j,z) - \partial_h^2\psi(x_j,z) + \partial_h^2(\pi_h\psi)(x_j,z)\right] \end{aligned}.$$

For any function ρ, the function $\partial_h^2\rho$ evaluated at nodes x_j of the grid is obtained as a linear combination of values of ρ also evaluated at the nodes of the grid:

$$\partial_h^2\rho(x_j) = \frac{1}{h^2}\sum_{p=1}^{n}\sum_{k=1}^{n}\nu_p\nu_k\left[\rho(x_{j+p+k-1}) - \rho(x_{j+p-k}) - \rho(x_{j-p+k}) + \rho(x_{j-p-k+1})\right]\ .$$

Since ψ and its interpolate $\pi_h\psi$ coincide at this nodes, the difference $\partial_h^2(\pi_h\psi)(x_j,z) - \partial_h^2\psi(x_j,z)$ vanishes and we get

$$E_h^{class} = c\left[-\frac{\partial^2\psi}{\partial x^2}(x_j,z) + \partial_h^2\psi(x_j,z)\right]\ .$$

Applying lemma 4.1 to the regular function ψ, we know that

$$\partial_h^2\psi(x_j,z) = \frac{\partial^2\psi}{\partial x^2}(x_j,z) + 2R_S^{[2n]}h^{2n}\frac{\partial^{2n+2}\psi}{\partial x^{2n+2}}(x_j,z) + O(h^{2n+2})\ ,$$

thus

$$E_h^{class} = 2cR_S^{[2n]}h^{2n}\frac{\partial^{2n+2}\psi}{\partial x^{2n+2}}(x_j,z) + O(h^{2n+2}) \ . \qquad \blacksquare$$

B Proof of Lemma 4.3

For a regular function ρ, the Taylor expansion of $I_\varepsilon \rho$ is

$$I_\varepsilon\rho(x) = \rho(x) + \sum_{k=1}^{n}\frac{\varepsilon^{2k}}{(2k)!}\rho^{(2k)}(x) + O(\varepsilon^{2n+2})$$

$$\Longrightarrow \quad \delta_h\rho(x) = \sum_{p=1}^{n}\mu_p\rho(x) + \sum_{k=1}^{n}\frac{h^{2k}}{2^{2k}\,(2k)!}\rho^{(2k)}(x)\sum_{p=1}^{n}\mu_p(2p-1)^{2k} + O(h^{2n+2}) \ .$$

Under the conditions (45), it remains

$$\delta_h\rho(x) = \rho(x) + \frac{h^{2n}}{2^{2n}\,(2n)!}\rho^{(2n)}(x)\sum_{p=1}^{n}\mu_p(2p-1)^{2n} + O(h^{2n+2}) \qquad \blacksquare$$

C Proof of proposition 4.1

Using equation (20)-(b) we rewrite the truncation error as

$$\begin{aligned} E_h^{mod} &= \omega^2\left(-\frac{1}{c}\varphi(x_j,z) + \sum_i (M_\alpha^{[2n]})_{ji}\varphi(x_i,z)\right) \\ &\quad -c\frac{\partial^2}{\partial x^2}\left(a\varphi + bw\right)(x_j,z) - \sum_i (K_h^{[2n]})_{ji}\left(a\varphi+bw\right)(x_i,z) \end{aligned} \ .$$

In the homogeneous case, we can evaluate the matrix $U_h^{[2n]}$ in the same way as we have evaluated $K_h^{[2n]}$ in the proof of lemma 4.2 (we omit in the following the index $[2n]$)

$$(U_h^{[2n]})_{ji} = \frac{1}{c}(\delta_h\chi_i,\delta_h\chi_j) = \frac{1}{c}(\delta_h^\star\delta_h\chi_i,\chi_j) = \frac{1}{c}(\delta_h^2\chi_i,\chi_j) \ .$$

Since $\delta_h^2\chi_i \in H_h$, it can be decomposed on the basis, thus

$$\delta_h^2\chi_i = \sqrt{h}\sum_m \delta_h^2\chi_i(x_m)\chi_m \Longrightarrow (U_h^{[2n]})_{ji} = \frac{1}{c}\sqrt{h}\delta_h^2\chi_i(x_j)$$

$$\Longrightarrow \quad \sum_i (U_h^{[2n]})_{ji}\varphi(x_i,z) = \frac{1}{c}\sqrt{h}\sum_i \delta_h^2\chi_i(x_j)\varphi(x_i,z) = \frac{1}{c}\delta_h^2(\pi_h\varphi)(x_j,z) \ .$$

From this expression, we deduce

$$\sum_i (M_\alpha^{[2n]})_{ji}\varphi(x_i,z) = \frac{1}{c}\left(\alpha\varphi(x_j,z) + (1-\alpha)\delta_h^2(\pi_h\varphi)(x_j,z)\right) \ ,$$

which leads to

$$\begin{aligned} E_h^{mod} &= \frac{\omega^2}{c}(1-\alpha)\left(\delta_h^2(\pi_h\varphi)(x_j,z) - \varphi(x_j,z)\right) + E_h^{class} \\ &= \frac{\omega^2}{c}(1-\alpha)\left(\delta_h^2(\pi_h\varphi)(x_j,z) - \delta_h^2\varphi(x_j,z) + \delta_h^2\varphi(x_j,z) - \varphi(x_j,z)\right) + E_h^{class} \end{aligned} \ .$$

Here again, for any function ρ, the values $\delta_h^2\rho(x_j)$ result from a linear combination of values of ρ at the grid points. Since φ and $\pi_h\varphi$ coincide at these grid points, the first difference in E_h^{mod} vanishes ant it remains

$$E_h^{mod} = \frac{\omega^2}{c}(1-\alpha)\left(\delta_h^2\varphi(x_j,z) - \varphi)(x_j,z)\right) + E_h^{class} \ ,$$

which is at least of order $2n$. The aim now is to determine $\alpha \in [0,1]$ such that the $2n$-order term also vanishes. Recall that

$$E_h^{class} = 2cR_S^{[2n]} h^{2n} \frac{\partial^{2n+2}\psi}{\partial x^{2n+2}}(x_j, z) + O(h^{2n+2}) \ .$$

On the other hand, we apply lemma 4.3 to the regular function φ

$$\delta_h^2 \varphi(x_j, z) - \varphi(x_j, z) = 2R_M^{[2n]} h^{2n} \frac{\partial^{2n}\varphi}{\partial x^{2n}}(x_j, z) + O(h^{2n+2}) \ .$$

These two results together give

$$\begin{aligned} E_h^{mod} &= \frac{\omega^2}{c}(1-\alpha)2R_M^{[2n]} h^{2n} \frac{\partial^{2n}\varphi}{\partial x^{2n}}(x_j, z) + 2cR_S^{[2n]} h^{2n} \frac{\partial^{2n+2}\psi}{\partial x^{2n+2}}(x_j, z) + O(h^{2n+2}) \\ &= 2h^{2n} \frac{\partial^{2n}}{\partial x^{2n}} \left((1-\alpha) R_M^{[2n]} \frac{\omega^2}{c}\varphi + cR_S^{[2n]} \frac{\partial^2 \psi}{\partial x^2} \right)(x_j, z) + O(h^{2n+2}) \end{aligned} .$$

One recognizes, between the brackets, equation (20)-(b) for an appropriate choice of α

$$(1-\alpha) R_M^{[2n]} = R_S^{[2n]} \quad \Longleftrightarrow \quad \alpha = \alpha^{[2n]} = 1 - \frac{R_S^{[2n]}}{R_M^{[2n]}} \ ,$$

thus for this value the truncation error becomes of order $2n+2$. We explicite the ratio by means of relation (53)

$$\frac{R_S^{[2n]}}{R_M^{[2n]}} = \frac{1}{2n+1} \frac{\displaystyle\sum_{p=1}^{n} \nu_p (2p-1)^{2n+1}}{\displaystyle\sum_{p=1}^{n} \mu_p (2p-1)^{2n}} = \frac{1}{2n+1} \quad \Longrightarrow \quad \alpha^{[2n]} = \frac{2n}{2n+1} \quad \in [0,1] \qquad \blacksquare$$

D Proof of proposition 4.2

• The first point is copied from the proof in the continuous case. Again, the uniqueness of the solution follows from the energy conservation, hence the problem is well-posed (existence and uniqueness are equivalent in a finite dimensional space).

• For the classical schemes, $A_\alpha^{[2n]} = K_h^{[2n]}$ and the condition (57) is obviously satisfied ($K_h^{[2n]} V_h \bullet V_h = \displaystyle\int_\Omega c|\partial_h v_h|^2 dx$).

• For the third point, we have to prove condition (57) for $0 \le \alpha \le 1$ in a homogeneous medium. We rewrite the stiffness matrix as

$$A_\alpha^{[2n]} = (I + (1-\alpha)R_h)^{-1} K_h^{[2n]} \ ,$$

with $R_h = (U_h^{[2n]} - M_h) M_h^{-1}$ and consider the quantity

$$\Upsilon = A_\alpha^{[2n]} V_h \bullet V_h = (I + (1-\alpha)R_h)^{-1} K_h^{[2n]} V_h \bullet V_h = K_h^{[2n]} V_h \bullet \left((I + (1-\alpha)R_h)^{-1} \right)^\star V_h \ .$$

We set $T_h = \left((I + (1-\alpha)R_h)^{-1} \right)^\star V_h$,

$$\Upsilon = K_h^{[2n]} (I + (1-\alpha)R_h)^\star T_h \bullet T_h = K_h^{[2n]} T_h \bullet T_h + (1-\alpha) K_h^{[2n]} R_h^\star T_h \bullet T_h \ .$$

The first term is a positive real number, it remains to examine the second term $\mathcal{T} = K_h^{[2n]} R_h^\star T_h \bullet T_h$. In the homeogeneous case, one can identify $M_h T_h$ with the function $\frac{1}{c} t_h$, $U_h T_h$ with $\frac{1}{c}\delta_h^2 t_h$, $K_h T_h$ with $-c\partial_h^2 t_h$ and thus rewrite

$$\mathcal{T} = c(\partial_h \delta_h^2 t_h, \partial_h t_h)_0 - c(\partial_h t_h, \partial_h t_h)_0 \ .$$

The main point now to conclude is to notice that operators δ_h and ∂_h commute, we thus have $\partial_h \delta_h^2 t_h = \delta_h^2 \partial_h t_h$ which yields

$$\mathcal{T} = c \parallel \delta_h \partial_h t_h \parallel_0^2 - c \parallel \partial_h t_h \parallel_0^2 \Longrightarrow \Upsilon = c \left(\alpha \parallel \partial_h t_h \parallel_0^2 + (1-\alpha) \parallel \delta_h \partial_h t_h \parallel_0^2 \right) .$$

The quantity Υ is therefore a positive real number. In the heterogeneous case, the difficulty comes from the fact that instead of $c(\partial_h \delta_h^2 t_h, \partial_h t_h)_0$ we end up with a quantity on the form $(c\partial_h c \delta_h \frac{1}{c} \delta_h t_h, \partial_h t_h)_0$ and nothing commute anymore! ■

E Numerical dispersion relation

We first rewrite (67) in the more convenient form :

$$\theta_e^{num} = \arctan \left(\frac{p_a}{1 - \frac{2}{\zeta_z} \arctan(\Sigma_K)} \right) \quad \text{with} \quad \Sigma_K = \frac{\Im m(N_K(\Delta z \widehat{C}_h))}{\Re e(N_K(\Delta z \widehat{C}_h))} . \tag{74}$$

E.1 second and fourth-order discretization in z

This corresponds to $K = 1$ and $K = 2$, for which we have the following Padé approximations :

$$N_1(x) = 1 + ix/2 \ ; \ N_2(x) = 1 + ix/2 - x^2/12 , \tag{75}$$

which give

$$\Sigma_1 = \frac{1}{2} \Delta z \widehat{C}_h \quad \text{and} \quad \Sigma_2 = \frac{\frac{1}{2} \Delta z \widehat{C}_h}{1 - \frac{1}{12} \Delta z^2 \widehat{C}_h^2} . \tag{76}$$

Remark E.1 We can check, using Taylor expansions, that

$$1 - \left(\frac{c k_z}{\omega} \right)^{num} = \widetilde{C}_h + O(\Delta z^{2K})$$

E.2 Discretization in the lateral variable

For the classical as well as for the modified discretizations, so long as α is not fixed, $\widehat{C}_h$ can be expressed with respect to the symbols $\widehat{M}_h \equiv \frac{1}{c}$, $\widehat{M}_\alpha^{[2n]}$, and $\widehat{K}_h^{[2n]}$, as

$$\widehat{C}_h = \frac{\omega}{c} \frac{b \widehat{K}_h^{[2n]}}{\omega^2 \widehat{M}_\alpha^{[2n]} - a \widehat{K}_h^{[2n]}} = \frac{\zeta_z}{\Delta z} \frac{b \frac{c}{\omega^2} \widehat{K}_h^{[2n]}}{c \widehat{M}_\alpha^{[2n]} - a \frac{c}{\omega^2} \widehat{K}_h^{[2n]}} .$$

Actually, it is more convenient to work with the quantities $\widetilde{K}_h^{[2n]} = \frac{c}{\omega^2} \widehat{K}_h^{[2n]}$, $\widetilde{M}_\alpha^{[2n]} = c \widehat{M}_\alpha^{[2n]}$ which can be expressed in terms of ζ and p_a, so we get

$$\Delta z \widehat{C}_{h;\, mod}^{[2n+2]} = \zeta_z \frac{b \widetilde{K}_h^{[2n]}}{\widetilde{M}_\alpha^{[2n]} - a \widetilde{K}_h^{[2n]}} \equiv \zeta_z \widetilde{C}_h ,$$

which gives (68) for respectively $\alpha = 1$ and $\alpha = \alpha^{[2n]}$.

Remark E.2 • It can be easily proved, applying the stiffness matrix to a plane wave and using the analysis of the order of the scheme (see proof of lemma 4.2) that

$$\widetilde{K}_h^{[2n]} = p_a^2 (1 + 2(-1)^n k_x^{2n} R_S^{[2n]} h^{2n} + O(h^{2n+2})) ,$$

which yields for the classical scheme

$$\widetilde{C}^{[2n]}_{h;\,class} = \frac{bp_a^2}{1-ap_a^2} + O(h^{2n}) = 1 - \left(\frac{ck_z}{\omega}\right)^{cont} + O(h^{2n})\ .$$

- In the same way, one gets for the modified mass matrix (see proof of proposition 4.1):

$$\widetilde{M}^{[2n]}_{\alpha} = 1 + 2(1-\alpha)(-1)^n k_x^{2n} R_M^{[2n]} h^{2n} + O(h^{2n+2})\ ,$$

thus

$$\widetilde{M}^{[2n]}_{\alpha} - a\widetilde{K}^{[2n]}_{h} = 1 - ap_a^2 + 2(-1)^n k_x^{2n} h^{2n}((1-\alpha)R_M^{[2n]} - ap_a^2 R_S^{[2n]}) + O(h^{2n+2})\ .$$

The optimal value $\alpha = \alpha^{[2n]}$ has been chosen such that $(1-\alpha^{[2n]})R_M^{[2n]} = R_S^{[2n]}$ so that

$$\widetilde{M}^{[2n]}_{\alpha} - a\widetilde{K}^{[2n]}_{h} = (1-ap_a^2)(1 + 2(-1)^n k_x^{2n} h^{2n} R_S^{[2n]} + O(h^{2n+2}))\ .$$

As could be expected, this naturally leads to a symbol of order $2n+2$

$$\widetilde{C}^{[2n+2]}_{h;\,mod} = \frac{bp_a^2}{1-ap_a^2} + O(h^{2n+2}) = 1 - \left(\frac{ck_z}{\omega}\right)^{cont} + O(h^{2n+2})\ .$$

We now present the expressions necessary for the classical and the modified schemes up to the order 6. From (68), we see that the classical schemes of order 2, 4 and 6 require the expressions of $\widetilde{K}^{[2]}_h$, $\widetilde{K}^{[4]}_h$ and $\widetilde{K}^{[6]}_h$ while the modified scheme of order 4 (respectively 6) requires the expressions of $\widetilde{K}^{[2]}_h$ and $\widetilde{M}^{[2]}_\alpha$ (respectively $\widetilde{K}^{[4]}_h$ and $\widetilde{M}^{[4]}_\alpha$)).

- **Symbol for the stiffness matrix**

$$\left\{\begin{aligned}
&\widetilde{K}^{[2]}_h(\zeta,p_a) = \frac{4}{\zeta^2}\sin^2(\frac{\zeta p_a}{2})\\
&\widetilde{K}^{[4]}_h(\zeta,p_a) = \frac{1}{\zeta^2}\left(\frac{9}{4}\sin(\frac{\zeta p_a}{2}) - \frac{1}{12}\sin(\frac{3\zeta p_a}{2})\right)^2\\
&\widetilde{K}^{[6]}_h(\zeta,p_a) = \frac{1}{\zeta^2}\left(\frac{75}{32}\sin(\frac{\zeta p_a}{2}) - \frac{25}{192}\sin(\frac{3\zeta p_a}{2}) + \frac{3}{320}\sin(\frac{5\zeta p_a}{2})\right)^2
\end{aligned}\right. . \tag{77}$$

- **Symbol for the mass matrix**

$$\left\{\begin{aligned}
&\widetilde{M}^{[2]}_\alpha(\zeta,p_a) = \frac{2}{3} + \frac{1}{3}\cos^2(\frac{\zeta p_a}{2}) && (\alpha^{[2]} = 2/3)\\
&\widetilde{M}^{[4]}_\alpha(\zeta,p_a) = \frac{4}{5} + \frac{1}{5}\left(\frac{9}{8}\cos(\frac{\zeta p_a}{2}) - \frac{1}{8}\cos(\frac{3\zeta p_a}{2})\right)^2 && (\alpha^{[4]} = 4/5)
\end{aligned}\right. . \tag{78}$$

References

[1] Georgios D. Akrivis and Vassilios A. Dougalis. On a conservative, high-order accurate finite element scheme for the "parabolic" equation. In *Computational acoustics, Vol. 1 (Princeton, NJ, 1989)*, pages 17–26. North-Holland, Amsterdam, 1990.

[2] Georgios D. Akrivis, Vassilios A. Dougalis, and Nikolaos A. Kampanis. Error estimates for finite element methods for a wide-angle parabolic equation. *Appl. Numer. Math.*, 16(1-2):81–100, 1994. A Festschrift to honor Professor Robert Vichnevetsky on his 65th birthday.

[3] A. Bamberger, B. Engquist, L. Halpern, and P. Joly. Higher order paraxial approximations for the wave equation. *SIAM J: Appl. Math.*, pages 128–154, 1988.

[4] A. Bamberger, B. Engquist, L. Halpern, and P. Joly. Paraxial approximations in heterogeneous media. *SIAM J: Appl. Math.*, pages 98–128, 1988.

[5] E. Bécache, F. Collino, and P. Joly. Some new developments about paraxial approximations for the 3d wave equation. Yearly report, PSI Consortium, 1994.

[6] E. Bécache, F. Collino, M. Kern, and Patrick Joly. On two migration methods based on paraxial equations in a 3d heterogeneous medium. In S. Hassanzadeh, editor, *Mathematical Methods in Geophysical Imaging III*. SPIE, 1995.

[7] J. P. Bérenger. A Perfectly Matched Layer for the Absorption of Electromagnetic Waves. *J. of Comp. Phys.*, 114:185–200, 1994.

[8] D. L. Brown. Applications of operator separation in reflection seismology. *Geophysics*, 3:288–294, 1983.

[9] H. Brysk. Numerical analysis of the 45-degree finite difference equation for migration. *Geophysics*, 48(5):532–542, 1983.

[10] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. North Holland, 1978.

[11] J. F. Claerbout. *Imaging the earth's interior*. Blackwell Scientific Publication co., 1983.

[12] F. Collino. Numerical analysis of mathematical models for wave propagation. Report, PSI Consortium, IFP, Rueil-Malmaison, France, 1993.

[13] F. Collino. Perfectly Matched Absorbing Layers for the Paraxial Equations . *J. of Comp. Phys.*, 131(1):164–180, 1997.

[14] F. Collino and P. Joly. Splitting of operators, alternate directions and paraxial approximations for the 3-D wave equation. *SIAM J. on Scientific and Stat. Comp.*, September, 1995.

[15] M. D. Collins. The time-domain solution of the wide-angle parabolic equation including the effects pf sediment dispersion. *J. Acoust. Soc. Am.*, 84 (6):2114–2125, December , 1988.

[16] A. A. Dubrulle. On numerical method for migration in layered media. *Geophysical Prospecting*, 56(11):237–264, 1983.

[17] D.M. Eidus. The principle of limiting absorption. *Amer. Math. Soc. Transl.*, 47:157–191, 1965.

[18] P. Froideveaux. First result of a 3-d prestack migration program. Annual Report, PSI Consortium, IFP, Rueil-Malmaison, France, 1990.

[19] R. W. Graves and R. W. Clayton. Modeling acoustic waves with paraxial extrapolators. *Geophysics*, 55:306–319, 1990.

[20] E. Hairer and G. Wanner. *Solving ordinary differential equations : 2 : stiff and differential-algebraic problems*. Springer Verlag, 1991.

[21] D. Hale. 3-d depth migration via mcclellan transformations. *Geophysics*, 11:1778–1785, november 1991.

[22] D. Hale. Stable explicit depth extrapolation of seismic wavefield. *Geophysics*, 56(11):1770–1777, november 1991.

[23] O. Holberg. Towards optimum one-way wave propagation. *Geophysical Prospecting*, 36:99–114, 1988.

[24] P. Joly and M. Kern. Numerical methods for 3-d migration. Annual Report, PSI Consortium, IFP, Rueil-Malmaison, France, 1990.

[25] M. Kern. Numerical methods for the 3-d paraxial wave equation in heterogeneous media,. Annual Report, PSI Consortium, IFP, Rueil-Malmaison, France, 1991.

[26] M. Kern. Numerical methods for the 3-d paraxial wave equation in heterogeneous media - part iii,. Annual Report, PSI Consortium, IFP, Rueil-Malmaison, France, 1992.

[27] M. Kern. A nonsplit 3d migration algorithm. In S. Hassanzadeh, editor, *Mathematical Methods in Geophysical Imaging III*. SPIE, 1995.

[28] D. Lee and A. D. Pierce. Parabolic equation development in recent decade. *J. of Comp. Acoustics*, 3(2):95–173, June 1995.

[29] Zhiming Li. Compensating finite-diffences errors in 3-d migration and modelling. *Geophysics*, 56(10):1650–1660, october 1991.

[30] E. L. Lindmann. Free-space boundary conditions for the time dependant wave equation. *Jour.Comp. Phys.*, 18, 1975.

[31] J. L. Lions and E. Magenes. *Problèmes aux limites non Homogènes et Applications*, volume 1. Dunod, 1968,.

[32] Z. Ma. Finite-difference migration with higher order approximation. *Oil Geophys. Prosp.*, 17:6–15, 1982.

[33] G. I. Marchuk. Splitting and alternating direction methods. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis*, volume I (Finite Difference Methods). Elsevier Science Publishers B.V. (North–Holland), Amsterdam, 1994.

[34] B. J. Orchard, W. L. Siegmann, and M. J. Jacobson. Three-dimensional time-domain paraxial approximations for ocean acoustic wave propagation. *J. Acoust. Soc. Am.*, 91 (2):788–801, February , 1992.

[35] D. Ristow and T.D. Ruhl. Fourier finite-difference migration. *Geophysics*, 59(12):1782–1893, December 1994.

[36] D. Ristow and T.D. Ruhl. 3-d implicit finite-difference migration by multiway splitting. *Geophysics*, 62(2):554–567, March 1997.

[37] F.D. Tappert. The parabolic approximation method. In *Wave propagation and underwater acoustics*, Lecture Notes in Physics, Vol. 70. J.B. Keller and J. Papadakis, eds., 1977.

[38] N. Tordjman. *Eléments finis d'ordre élevé avec condensation de masse pour l'équation des ondes.* PhD thesis, Univ. Paris IX, 1995.

Unit´e de recherche INRIA Lorraine, Technopôle de Nancy-Brabois, Campus scientifique,
615 rue du Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unit´e de recherche INRIA Rennes, Irisa, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unit´e de recherche INRIA Rhône-Alpes, 655, avenue de l'Europe, 38330 MONTBONNOT ST MARTIN
Unit´e de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
Unit´e de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex (France)
http://www.inria.fr
ISSN 0249-6399